



Tesis doctoral

Relación entre el cambio basado en los centros y el cambio basado en los individuos

Eduardo Estrada Alonso
Madrid, mayo de 2015



Departamento de Psicología Social y Metodología
Facultad de Psicología

Tesis doctoral
Programa de doctorado en
Metodología de las Ciencias del Comportamiento

Relación entre el cambio basado en los centros y el cambio basado en los individuos

Doctorando: Eduardo Estrada Alonso
Director: Antonio Pardo Merino

Agradecimientos

Esta tesis doctoral nunca habría existido de no ser por un conjunto de personas excepcionales que se han ido cruzando en mi camino. Me refiero a todos los maestros y maestras que he tenido en distintos momentos de mi vida que, en lugar de conformarse con hacer sólo lo imprescindible para “cubrir el expediente”, fueron mucho más allá e invirtieron (e invierten) su tiempo, su energía y su cariño en hacer las cosas bien.

En realidad considero como maestro/a a todo aquel con la generosidad suficiente como para empeñarse en ayudar a los demás a ser mejores, lo cual invariablemente requiere de dosis masivas de paciencia y pedagogía. Pero dado que este es un trabajo académico, agradezco aquí especialmente todas las horas de más preparando las clases del día siguiente, todas las tutorías, conversaciones y reuniones que acabaron alargándose mucho más de lo razonable, todas las búsquedas de referencias bibliográficas para otros. Ahora que yo mismo me veo en situaciones parecidas, entiendo mejor el valor y el coste de cada uno de esos pequeños derroches. Algunas de esas personas leerán esto, otras no. Pero cada una de ellas ha provocado pequeños giros en mi trayectoria que me han acabado trayendo hasta aquí. Sólo gracias a esos cruces de caminos se entiende que, partiendo desde donde partí, haya llegado hasta donde estoy ahora.

De todos esos maestros, el más importante ha sido mi director de tesis. Antonio es una inspiración constante, y el espejo en el que me miro para valorar si voy consiguiendo ser mejor científico, mejor filósofo y mejor profesor. Trabajar juntos ha influido decisivamente en la manera en que comprendo la ciencia y el mundo. Pero, al margen de mi formación académica, nuestra colaboración me ha

permitido aprender muchas otras cosas de las que ni siquiera soy totalmente consciente, desde mi manera de escribir hasta cómo contar chistes. He aprendido mucho de él desde el punto de vista científico, pero quizá haya aprendido todavía más desde el punto de vista personal. Mi concepción de lo que supone ser una buena persona (en el sentido más amplio de la palabra) está afectada por esto, y mi progresión a partir de ahora como académico y como ser humano, me lleve a donde me lleve, estará marcada por lo que he aprendido de él. Por ello, todo agradecimiento que pueda expresarle en estas páginas es insuficiente.

Esta tesis tampoco existiría sin el apoyo constante e incondicional de mis padres. Soy plenamente consciente de que llevan toda la vida poniendo todo de su parte, de que nunca han escatimado esfuerzos ni energías, para que mi hermano y yo tengamos todas las oportunidades posibles. Si he llegado hasta este punto es, sobre todo, porque lo único que he tenido que hacer es ir aprovechando esas oportunidades. Por eso, va para ellos todo mi reconocimiento, gratitud y cariño.

También le debo mucho la persona con la que comparto mi vida; Nuria. Acabaría antes enumerando las cosas por las que *no* le estoy agradecido que las cosas por las que sí (literalmente). Desde escribir una tesis doctoral hasta sobrellevar los días malos, todo es más fácil a su lado. La luz y la alegría que irradia me mueven a seguir adelante. Su pensamiento lúcido afila el mío. De hecho, es suya una de las ideas centrales de la discusión de este trabajo. Es el mayor regalo que la vida me ha dado, y su huella está en todo lo que hago.

Desde que nos conocemos, el profesor Emilio Ferrer me ha tratado excepcionalmente bien. Además ser el principal culpable de que mi estancia en UC Davis fuera enriquecedora desde todos los puntos de vista, me ha ayudado repetida y generosamente en varios momentos de mi corta carrera académica. El último ejemplo de ello es que se haya prestado de forma altruista a valorar esta

tesis para que pudiera obtener la Mención Internacional. Esto último se lo agradezco también profesor Rodrigo Ferrer.

La contribución de varias personas sido muy importante para mejorar el presente trabajo. La factura técnica de la impresión en papel es consecuencia del buen hacer y la profesionalidad de Manuel Carmona. Juan Miguel Sánchez ha tenido la amabilidad de revisar las secciones en inglés. María, Aline y Adrien (además de Nuria, a la que ya he mencionado) se implicaron decididamente en la búsqueda de una imagen para la portada, alegrándome además una bonita tarde de primavera.

Índice de contenidos

Brief Introduction and Abstract	11
Abstract	13
Introducción	17
Resumen	19
Antecedentes	21
La evaluación del cambio basada en los centros	25
La evaluación del cambio basada en los individuos	32
Objetivos	39
Estudio 1	43
El caso de los diseños pre-post de grupo único	
Objetivo	43
Método	44
Resultados	51
Discusión	61

Estudio 2	67
El caso de los diseños de grupos aleatorios	
Objetivo	69
Método	69
Resultados	73
Discusión	80
 Estudio 3	 85
El caso de los diseños pre-post con grupos aleatorios	
Objetivo	86
Método	87
Resultados	93
Discusión	107
 Discusión general	 111
 General Discussion	 117
 Referencias bibliográficas	 123
Apéndice 1. Tablas de equivalencias entre métodos basados en los centros y basados en los individuos	131
Apéndice 2. Código de MatLab utilizado en las simulaciones	139

Brief Introduction and Abstract

In various areas of psychology, and particularly in the clinical field, it is common to perform interventions in order to cause a change. To evaluate whether an intervention had an effect, it is necessary to assess when the change has occurred.

There are many different methods to evaluate when a change happened (see for example Ogles, Lunnen, & Bonesteel, 2001). In addition to their own judgment, therapists may take into consideration the opinions provided by informants close to the treated person (Achenbach, 2006; De Los Reyes, Thomas, Goodman, & Kundey, 2013; Oltmanns & Turkheimer, 2009), objective indicators such as medical tests or behavior frequency counts (Kahn, Goldfarb, Pollack, & Peck, 1960; Morris et al., 2012; Owens et al., 2009), etc.

Change can also be evaluated by using the person's responses as a reference. This is a frequent strategy in psychology, in which there is a long tradition of measuring through standardized tests, scales and questionnaires. This method has also become popular in the last 20 years in the medical field to assess, among others, the health related quality of life (*HRQOL*) (Fayers & Machin, 2000; Lemieux,

Beaton, Hogg-Johnson, Bordeleau, & Goodwin, 2007; Norman, Sridhar, Guyatt, & Walter, 2001) and treatment satisfaction (Aharony & Strasser, 1993; Albrecht & Hoogstraten, 1998; Cleary & McNeil, 1988; Shikar & Rentz, 2004; Wright, 2000).

To evaluate whether a particular score on a scale or questionnaire indicates a change, two different approaches can be adopted: one is based on the *centers of the scores' distributions* and the other is based on *the individual scores*. We will refer to these approaches as *centers-based* and *individuals-based*.

In the **centers-based approach** the scope is to find out whether the treated group (as a whole) has undergone a change (*group change*). The usual way to assess the group change is by comparing the pre and post means (in a pre-post design) or the means of the experimental and control groups (in a design with randomized groups) through *hypothesis testing* and *effect size measures*. The statistical significance yielded by the hypothesis tests indicate whether the compared means differ beyond the variability attributable to sampling random fluctuations; the effect size measures is used to quantify the magnitude of those differences.

In the **individuals-based approach** the scope is to find out which specific individuals experienced a change (*individual change*). The usual way to assess individual change is by applying various statistics called *reliable change indices*. Some of these indices are based on the standardization of the pre-post differences, some on the standard error of measurement, and some others on the linear regression predictions (Crosby, Kolotkin, & Williams, 2003; Ferrer & Pardo, 2014); but all of them are designed to identify which specific individuals change. The advantage of this approach over the centers-based alternative is that, in addition to knowing which individuals change, it is also possible to know the percentage of subjects that change and, through that percentage, the change experienced by the

group as a whole. Therefore, the individuals-based approach can provide information about both the individual change and the group change.

In principle, one could think that the centers-based approach and the individuals-based approach offer very different information. However, as we shall see, this is not exactly true. In fact, some of the information they offer is very similar. The results of Norman et al. (2001) suggest that, at least under certain conditions, it is possible to predict some information provided by one approach from the information provided by the other.

The **overall objective** of this work is to study the possible relationship between these two alternative ways of assessing change, in order to evaluate to what extent they provide the same information. The motivation to study this relationship is not only to find possible matches on the information obtained with both approaches, but also to find out whether the specific information from each approach contributes to enrich the information provided by the other.

Abstract

This work consists of a **review** and **three studies**. The **review** focused on the available centers-based and individuals-based strategies for assessing change. As a result of this review we had to make two decisions that affected the whole work. First, among the many available statistics for quantifying the group change (centers-based approach) we decided to use some measures of effect size; for quantifying individual change (individuals-based approach) we chose some reliable change indices. Second, given that the specific way to assess the change depends decisively on the research design, we had to choose only some designs (we chose the three most common in the field of health sciences) and study the relationship between the two approaches separately for each design.

In the **first study** we used a *single group pre-post* design. As a centers-based measure of change we used Cohen's standardized difference d ; as an individuals-based measure of change we used the percentage of changes (obtained from individual changes identified by the *standardized individual difference* and the *reliable change index*). After studying the relationship between the two approaches in different conditions, we found that (1) there is a relationship between the two approaches; (2) the relationship is linear and shows an excellent fit; (3) the relationship is consistently found across the various simulated conditions.

In the **second study** we used a *randomized groups* design. As a centers-based measure of change we used Cohen's standardized difference d ; as an individuals-based measure of change we used the statistic called *probability of superiority*. After studying the relationship between the two approaches in different conditions, we found that (1) there is a relationship between the two approaches; (2) the relationship is quadratic and shows an excellent fit; (3) the relationship is consistently found across the various simulated conditions.

In the **third study** we used a *randomized groups pre-post* design. As a centers-based measure of change we used the Hays' ω^2 effect size estimate corresponding to the *group x time* interaction effect; as an individuals-based measure of change we used the *net percentage of individual improvements*. After studying the relationship between the two approaches in different conditions, we found that (1) there is a relationship between the two approaches; (2) the relationship is linear and shows an excellent fit; (3) the relationship is reasonably consistent across the various simulated conditions.

The most interesting finding about the detected relationships is the fact that the information yielded by each approach helps completing and better understanding the information obtained with the other approach. Far from being

independent, our results lead us to think that both approaches are different expressions of essentially the same information. However, reporting the percentage changes (or a similar statistic) has the advantage of being easier to understand than a number expressed in standard deviation units.

Introducción

En diversos ámbitos de la Psicología, y muy especialmente en el ámbito clínico, es habitual realizar intervenciones con el objetivo de provocar algún cambio. Valorar si esas intervenciones tienen algún efecto requiere poder determinar cuándo se ha producido un cambio.

Existen distintos métodos para evaluar si se ha producido un cambio (Ogles, Lunnen, & Bonesteel, 2001). Además del criterio del terapeuta, puede tenerse en cuenta la información que aportan los familiares del sujeto (Achenbach, 2006; De Los Reyes, Thomas, Goodman, & Kundey, 2013; Oltmanns & Turkheimer, 2009), indicadores objetivos como pruebas médicas o recuentos de frecuencias de conductas (Kahn, Goldfarb, Pollack, & Peck, 1960; T. R. Morris et al., 2012; Owens et al., 2009), etc.

También se puede evaluar el cambio tomando como referencia las propias respuestas de los sujetos. Esta estrategia es muy utilizada en psicología, donde existe una larga tradición basada en medir mediante tests, escalas o cuestionarios; y ha tomado fuerza en las dos últimas décadas en el ámbito médico para valorar,

entre otras cosas, la calidad de vida relacionada con la salud (*health related quality of life* o *HRQOL*; Fayers & Machin, 2000; Lemieux, Beaton, Hogg-Johnson, Bordeleau, & Goodwin, 2007; Norman, Sridhar, Guyatt, & Walter, 2001), y la satisfacción con el tratamiento (Aharony & Strasser, 1993; Albrecht & Hoogstraten, 1998; Cleary & McNeil, 1988; Shikier & Rentz, 2004; Wright, 2000).

Para evaluar si una puntuación en una escala o cuestionario indica un cambio, se pueden adoptar dos enfoques distintos: el basado en los *centros de las distribuciones* y el basado en las *puntuaciones individuales*. Llamaremos estos dos enfoques el *basado en los centros* y el *basado en los individuos*.

En el **enfoque basado en los centros** se intenta averiguar si el grupo tratado (en su conjunto) ha experimentado un cambio (*cambio grupal*). La forma habitual de valorar el cambio grupal consiste en comparar las medias del pre y del post (si se trata de un diseño pre-post) o las medias de los grupos experimental y control (si se trata de un diseño con grupos aleatorios) mediante la aplicación de *contrastes de hipótesis y medidas del tamaño del efecto*. La significación estadística que ofrecen los contrastes de hipótesis indica si los promedios comparados difieren más allá de la variabilidad atribuible a las fluctuaciones aleatorias propias del azar muestral; las medidas del tamaño del efecto intentan cuantificar la magnitud de esa diferencia.

En el **enfoque basado en los individuos** se intenta averiguar qué individuos concretos han experimentado un cambio (*cambio individual*). La forma habitual de valorar el cambio individual consiste en aplicar diversos estadísticos denominados *índices de cambio fiable (reliable change indices)*. Algunos de estos índices están basados en la tipificación de las diferencias pre-post, otros en el error típico de medida, otros en los pronósticos de la regresión lineal (Crosby, Kolotkin, & Williams, 2003; Ferrer & Pardo, 2014); pero todos ellos han sido diseñados para

poder identificar qué individuos concretos cambian. La ventaja de este enfoque sobre el enfoque basado en los centros es que, aquí, además de saber qué individuos cambian, también es posible conocer el porcentaje de sujetos que cambian y, a través de ese porcentaje, el cambio experimentado por el grupo en su conjunto. Por tanto, el enfoque basado en los individuos, además de aportar información acerca del cambio individual, también aporta información sobre el cambio grupal.

En principio, podría pensarse que el enfoque basado en los centros y el enfoque basado en los individuos ofrecen información muy distinta. Sin embargo, según tendremos ocasión de comprobar, esto no es exactamente así. De hecho, parte de la información que ofrecen es muy parecida. Los resultados de Norman *et al.* (2001) sugieren que, al menos bajo ciertas condiciones, es posible obtener parte de la información que ofrece un enfoque a partir de la información que ofrece el otro.

El **objetivo general** de este trabajo es precisamente estudiar la posible relación existente entre estas dos formas alternativas de evaluar el cambio, con la intención de poder valorar en qué medida aportan o no la misma información. Y la motivación para estudiar esta relación está no solo en encontrar posibles coincidencias en la información que se obtiene con ambos enfoques, sino también en averiguar si la información específica de cada enfoque puede contribuir a enriquecer la información que aporta el otro.

Resumen

Este trabajo consta de **una revisión y tres estudios**. La **revisión** se ha centrado en las estrategias disponibles para valorar el cambio basado en los centros y el cambio basado en los individuos. Como consecuencia de esta revisión hemos

tenido que tomar dos decisiones que han afectado a todo el trabajo. En primer lugar, entre los múltiples estadísticos disponibles para cuantificar el cambio grupal (enfoque basado en los centros) hemos optado por utilizar algunas medidas del tamaño del efecto; para cuantificar el cambio individual (enfoque basado en los individuos) hemos optado por utilizar algunos índices de cambio fiable. En segundo lugar, dado que la forma concreta de valorar el cambio depende de manera determinante del diseño de investigación utilizado, hemos tenido que elegir solo algunos diseños (hemos elegido los tres más utilizados en el ámbito de las ciencias de la salud) y estudiar la relación entre ambos enfoques separadamente para cada diseño.

En el **primer estudio** hemos utilizado un diseño *pre-post de grupo único*. Como medida del cambio basada en los centros hemos utilizado la *diferencia estandarizada d* de Cohen; como medida del cambio basada en los individuos hemos utilizado el *porcentaje de cambios* (obtenido éste, a su vez, a partir de los cambios individuales identificados mediante la *diferencia individual tipificada* y el *índice de cambio fiable*). Tras estudiar la relación entre ambos enfoques en diferentes condiciones, hemos encontrado que: (1) existe relación entre ambos enfoques; (2) la relación es de tipo lineal y con un excelente grado de ajuste; (3) la relación encontrada es consistente a través de las diferentes condiciones simuladas.

En el **segundo estudio** hemos utilizado un diseño de *grupos aleatorios*. Como medida del cambio basada en los centros hemos utilizado la *diferencia estandarizada d* de Cohen; como medida del cambio basado en los individuos hemos utilizado el estadístico *probabilidad de superioridad*. Tras estudiar la relación entre ambos enfoques en diferentes condiciones, hemos encontrado que: (1) existe relación entre ambos enfoques; (2) la relación es de tipo cuadrática y con

un excelente grado de ajuste; (3) la relación encontrada es consistente a través de las diferentes condiciones simuladas.

En el **tercer estudio** hemos utilizado un diseño *pre-post con grupos aleatorios*. Como medida del cambio basado en los centros hemos utilizado la medida del tamaño del efecto ω^2 de Hays correspondiente al efecto de la interacción *grupo x momento*; como medida del cambio basado en los individuos hemos utilizado el *porcentaje neto de mejorías*. Tras estudiar la relación entre ambos enfoques en diferentes condiciones, hemos encontrado que: (1) existe relación entre ambos enfoques; (2) la relación es de tipo lineal y con un excelente grado de ajuste; (3) la relación encontrada es razonablemente consistente a través de las diferentes condiciones simuladas.

Lo interesante de las relaciones detectadas es que la información que se obtiene con cada enfoque ayuda a completar y entender mejor la información que se obtiene con el otro enfoque. Lejos de ser independientes, nuestros resultados nos llevan a pensar que ambos enfoques son expresiones distintas de prácticamente la misma información. Pero informar del porcentaje de cambios (o de estadísticos similares) tiene la ventaja de que es más fácil de entender que un número expresado en unidades de desviación típica.

Antecedentes

La primera idea que surge espontáneamente al reflexionar sobre el enfoque basado en los centros y el enfoque basado en los individuos es que la información que ofrece cada uno ha de ser necesariamente distinta. De hecho, la literatura científica sobre el tema sostiene una opinión aparentemente unánime en ese sentido.

Por ejemplo, Ottenbacher, Johnson y Hojem, al hablar de la comparación de grupos, señalan que “los resultados de estos estudios están basados en la mejoría de las puntuaciones en promedio para todos los sujetos y no proporcionan información sobre el comportamiento de los pacientes individuales” (Ottenbacher, Johnson, & Hojem, 1988, p. 115).

Testa manifiesta una opinión similar al afirmar que “el valor medio (...) es a menudo inadecuado para trasladar los efectos del tratamiento a planes de acción porque solo reflejan los riesgos o beneficios en el centro de la distribución” (Testa, 2000, pp. 170-171).

En la misma línea, Schimtt y Di Fabio afirman que “el cambio estadísticamente significativo a nivel grupal puede no ser significativo a nivel individual (...). Los cambios en la media de un grupo pueden ser resultado de unos pocos individuos con cambios relativamente grandes, o de muchos individuos con cambios relativamente pequeños” (Schmitt & Di Fabio, 2004, pp. 1008-1009).

Lambert, Hansen y Bauer, al hablar del efecto de un tratamiento para la reducción del sobrepeso, señalan que “los contrastes de hipótesis no ofrecen información sobre la variedad de respuestas al tratamiento dentro del grupo tratado. Con una pérdida de peso de 8 libras, algunos individuos pueden haber perdido 16 libras, mientras que otros pueden no haber perdido peso o incluso haberlo ganado” (Lambert, Hansen, & Bauer, 2008, pp. 359-360).

Algo parecido opinan Vindras, Desmurget y Baraduc, que sostienen que “el efecto de un factor puede ser significativo para cada individuo (en comparación con la variabilidad intra-individual), aunque que las pruebas de Student y Fisher ofrezcan valores p próximos a uno si el promedio poblacional es lo suficientemente pequeño” (Vindras, Desmurget, & Baraduc, 2012, p. 2).

Por tanto, en principio, parece claro que el cambio basado en los centros y el basado en las puntuaciones individuales son cosas bien *distintas*: es evidente que conocer cuánto cambian los centros no nos dice nada acerca de qué individuos concretos cambian.

Sin embargo, no parece que estos dos enfoques sean del todo *independientes*. Cuando el objetivo de un estudio es evaluar el efecto de una intervención, tanto el enfoque basado en los centros como el basado en los individuos proporcionan información útil. El cambio en los centros indica si se ha producido o no un cambio en el grupo (en el grupo tomado en su conjunto) y si ese cambio es *fiable* (significación estadística) o *relevante* (tamaño del efecto). El cambio basado en los individuos indica qué personas cambian y, además, permite calcular el porcentaje de cambios dentro del grupo tratado, lo cual constituye una forma distinta de evaluar la magnitud del efecto de la intervención.

En otras palabras, tanto el *cambio los centros* como el *porcentaje de cambios* son medidas grupales obtenidas a partir del cambio individual. En consecuencia, parece lógico esperar que el porcentaje de individuos que cambian no sea completamente independiente del cambio en el centro de la distribución.

En un diseño pre-post, por ejemplo, cabe esperar que cuanto más alejados estén los centros de las distribuciones del pre y del post, mayor sea el porcentaje de cambios individuales; es decir, cabe esperar que cuanto mayor sea la media de las diferencias pre-post, más probable sea que una determinada diferencia individual pre-post sobrepase el umbral establecido para decidir que el individuo ha cambiado. Por ejemplo, en una distribución normal, si la media de las diferencias vale cero, la probabilidad de encontrar casos por encima de 1.645 desviaciones típicas vale 0.05; si la media es 0.5 desviaciones típicas mayor que cero, esa probabilidad vale 0.13, etc. Por tanto, en condiciones de normalidad, el

porcentaje de cambios individuales es función directa de la magnitud del cambio en los centros. Sin embargo, en escenarios reales, donde la normalidad es más bien infrecuente (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Micceri, 1989), aunque las distribuciones del pre y del post suelen ser conocidas, la distribución de las diferencias no solo es desconocida sino que no es nada fácil deducirla a partir de las puntuaciones del pre y del post.

La no independencia entre el tamaño del efecto y el porcentaje de cambios ha sido sugerida, por ejemplo, por Norman *et al.* (2001), que encontraron que el tamaño del efecto observado tras una intervención (la magnitud del cambio en el centro de la distribución de las diferencias pre-post) es el principal determinante de la proporción de sujetos que se benefician de dicha intervención. Basándose en datos simulados, llegaron a la conclusión de que la relación entre el tamaño del efecto y el porcentaje de cambios es básicamente de tipo lineal para efectos de tamaño menor que uno (independientemente del punto de corte elegido para decidir que se ha producido un cambio, y con distribuciones normales y moderadamente asimétricas). Lemieux *et al.* (2007) han llegado a conclusiones parecidas utilizando datos reales en lugar de simulados.

El cambio en el centro de la distribución y el porcentaje de cambios individuales son, ambos, estadísticos *grupales*, pero se calculan de formas muy distintas: para calcular el *centro de la distribución* se utilizan todos los casos y cada uno contribuye con su parte proporcional de cambio; para calcular el *porcentaje de cambios* solo se tienen en cuenta los cambios que superan un determinado punto de corte y todos los cambios son ponderados por igual independientemente de su magnitud. Este dato podría hacer pensar que el cambio en el centro de la distribución, no solo no está informando sobre qué individuos cambian (cosa que parece concentrar un acuerdo razonable; Lambert, Hansen, & Bauer, 2008, pp.

359-360; Ottenbacher, Johnson, & Hojem, 1988, p. 115; Schmitt & Di Fabio, 2004, pp. 1008-1009; Testa, 2000, pp. 170-171; Vindras, Desmurget, & Baraduc, 2012, p. 2), sino que tampoco está informando de cuántos individuos cambian, pues un mismo cambio en el centro de la distribución (es decir, un cambio de determinada magnitud en la diferencia entre los centros) podría ser el resultado de unos pocos individuos que cambian mucho o el resultado de muchos individuos que cambian poco.

Sin embargo, los resultados de Norman *et al.* (2001) y de Lemieux *et al.* (2007) apuntan a que, bajo ciertas condiciones, el cambio en el centro de la distribución puede estar informando sobre el porcentaje de cambios individuales. Es decir, los resultados de Norman y sus colaboradores sugieren que, al menos bajo ciertas condiciones, es posible obtener parte de la información que ofrece un enfoque a partir de la información que ofrece el otro.

La evaluación del cambio basada en los centros

La forma concreta de valorar el cambio (tanto en el enfoque basado en los centros como en el basado en los individuos) viene determinada por el diseño de investigación utilizado. En este trabajo, para estudiar la relación entre la información que ofrecen ambos enfoques nos vamos a centrar en tres de los diseños más utilizados en las ciencias sociales y de la salud: el diseño *pre-post de grupo único*, el diseño de *grupos aleatorios* y el diseño *pre-post con grupos aleatorios*.

Diseño pre-post de grupo único

En un *diseño pre-post de grupo único*, también llamado *diseño con los mismos sujetos* en terminología de León y Montero (2003), o *diseño de medidas totalmente repetidas* en terminología de Ato, López y Benavente (2013), se toman dos medidas a los mismos individuos antes (pre) y después (post) de la intervención y el cambio se valora comparando los momentos pre y post.

La debilidad principal de este diseño es que no permite separar el efecto de la intervención del efecto de otros factores no controlados como, por ejemplo, la maduración debida al paso del tiempo. Sin embargo consideramos que es un buen punto de partida para nuestro estudio porque, además de ser un diseño que se utiliza con cierta frecuencia en la investigación aplicada, es el escenario en el que se han desarrollado los índices de cambio individual que estudiaremos más adelante (Crawford, Howell, & Garthwaite, 1998; Hageman & Arrindell, 1999; Jacobson & Truax, 1991; Payne & Jones, 1957; Wyrwich, Tierney, & Wolinsky, 1999).

La forma de valorar el cambio basado en los centros en un *diseño pre-post de grupo único* consiste simplemente en comparar la media del pre y la media del post (o, lo que es equivalente, en averiguar si la media de las diferencias pre-post es distinta de cero). Para decidir si las medias del pre y del post son distintas, es decir, para decidir si se ha producido o no un cambio en el centro de la distribución de las diferencias pre-post (significación estadística), suele utilizarse la *prueba T de Student para muestras relacionadas*; y para cuantificar la magnitud de ese cambio (tamaño del efecto) suele utilizarse¹ la *diferencia estandarizada d de Cohen* (1988).

¹ El grado de relación entre los momentos pre y post cuantificado mediante coeficientes de correlación como el de Pearson también es un indicador del tamaño del efecto, pero de *otro* efecto. En un diseño pre-post, la diferencia entre promedios (que es el estadístico que está informando del

Dado que el valor del estadístico T depende no solo del tamaño de la diferencia entre las medias del pre y del post sino del tamaño de la muestra elegida, la significación estadística no suele ser un criterio útil para valorar la magnitud del cambio observado. Para valorar la magnitud del cambio utilizaremos la media estandarizada d de Cohen. Esta consideración es extensible a todos los diseños con los que trabajaremos: en todos ellos utilizaremos una medida del tamaño del efecto como criterio para cuantificar la magnitud del cambio.

Lo que hace la media estandariza d de Cohen es expresar el valor absoluto de la diferencia entre la media del pre (M_{pre}) y la media del post (M_{post}) en unidades de desviación típica; o, lo que es equivalente, expresar en unidades de desviación típica la media de las diferencias pre-post ($M_{\text{dif}} = |M_{\text{pre}} - M_{\text{post}}|$). Por tanto, d no es más que una puntuación típica: la puntuación típica que corresponde al valor absoluto de la media de las diferencias pre-post.

Ahora bien, la tipificación de M_{dif} puede realizarse de dos maneras distintas: dividiéndola entre la *desviación típica del pre* (S_{pre}) o dividiéndola entre la *desviación típica de las diferencias pre-post* (S_{dif}). Por tanto, el estadístico d puede expresarse en dos métricas distintas: la del *pre* y la de las diferencias *pre-post*.

Algunos autores, incluyendo el propio Cohen, recomiendan utilizar la métrica del *pre* por considerar que la referencia natural para pensar sobre posibles cambios en las puntuaciones del pre debe ser la variabilidad original (la variabilidad del *pre*) y no la variabilidad en las diferencias *pre-post* (Cohen, 1969; Cumming & Finch, 2001). Además, como la desviación típica de las diferencias pre-post es tanto menor cuanto mayor es la correlación entre el pre y el *post*² algunos

cambio en el centro de la distribución de las diferencias pre-post) es del todo independiente del grado de relación existente entre el pre y el post.

² Puesto que $S_{\text{dif}} = (S_{\text{pre}}^2 + S_{\text{post}}^2 - 2S_{\text{pre-post}})^{1/2}$, cuanto mayor sea el valor de $S_{\text{pre-post}}$ (es decir, cuanto mayor sea el grado de relación pre-post), menor será el valor de S_{dif} .

autores sugieren utilizar la métrica de *pre* para evitar valores inflados de d (Cumming, 2012).

Sin embargo, en nuestro estudio hemos decidido estandarizar d utilizando la métrica de las diferencias, decir, $d = M_{\text{dif}} / S_{\text{dif}}$. La decisión de utilizar esta métrica obedece a dos motivos. En primer lugar, consideramos que un estadístico debe ser tipificado utilizando como referencia su propia distribución; por tanto, dado que el estadístico que se está tipificando es la media de las diferencias, lo razonable es utilizar como referencia la distribución de las diferencias y no la distribución de las puntuaciones del *pre*. En segundo lugar, los índices diseñados para evaluar el cambio individual (es decir, los índices que se utilizan para obtener el porcentaje de cambios que utilizaremos más adelante) se obtienen a partir de la distribución de las diferencias; dado que el objetivo de este trabajo es estudiar la relación entre ambos enfoques (el basado en el centro, y el basado en los individuos), lo razonable es utilizar estadísticos basados en la misma distribución: la distribución de las diferencias.

En cualquier caso, la métrica concreta del estadístico d , es decir, el hecho de que el estadístico d se obtenga en la métrica de las puntuaciones *pre* o en la métrica de las diferencias *pre-post*, no es relevante para nuestro trabajo. Si

$$d_1 = M_{\text{dif}} / S_{\text{dif}}$$

(es decir, si d_1 es la diferencia estandarizada d expresada en la métrica de las diferencias *pre-post*) y

$$d_2 = M_{\text{dif}} / S_{\text{pre}}$$

(es decir, la diferencia tipificada d expresada en la métrica de las puntuaciones del *pre*), entonces d_2 puede ser transformada fácilmente en d_1 mediante (Cohen, 1969, pp. 48-49):

$$d_1 = d_2 [2(1 - R_{XY})]^{1/2}$$

Y, según tendremos ocasión de comprobar más adelante, las conclusiones sobre d_1 son del todo válidas para d_2 .

Diseño de grupos aleatorios

Este diseño consiste en formar aleatoriamente dos grupos, aplicar la intervención solo a uno de ellos (grupo experimental o tratado) y a continuación medir a todos los participantes en la variable de interés (Kirk, 2013, p. 30). En este caso la intervención se evalúa mediante la comparación entre el grupo tratado y el no tratado (control).

De cara a la valoración de la intervención, este diseño tiene el inconveniente de que la falta de seguimiento impide valorar si el efecto estudiado perdura en el tiempo. Además, con este tipo de diseños siempre queda la duda de si los grupos formados son completamente equivalentes³; consecuentemente, no ofrecen plenas garantías de que las eventuales diferencias sean debidas exclusivamente a la intervención.

Para estimar el tamaño del efecto en un diseño de grupos aleatorios también suele utilizarse la diferencia estandarizada d de Cohen (1988):

$$d = \frac{M_{\text{tto}} - M_{\text{ctrl}}}{S_{\text{?}}}$$

La expresión $S_{\text{?}}$ es una referencia explícita a la falta de consenso existente sobre la forma correcta de estimar la desviación típica que debe tomarse como referencia

³ Existen algunos procedimientos que pueden ayudar a hacer una valoración del grado de equivalencia existente entre dos grupos: nos referimos los contrastes de equivalencia (Pardo & San Martín, 2015, pp. 191-197). Pero la información que se obtiene por esta vía es más bien limitada: no permite valorar la equivalencia en todas las variables que podrían ser fuente de diferenciación, sino solo en aquellas identificadas como tales.

para definir la métrica del estadístico d . No obstante, la expresión que ha recibido mayor consenso es la que se basa en la combinación de la variabilidad de las puntuaciones de ambos grupos:

$$S_d = \sqrt{\frac{(n_{\text{tto}} - 1) S_{\text{tto}}^2 + (n_{\text{ctrl}} - 1) S_{\text{ctrl}}^2}{n_{\text{tto}} + n_{\text{ctrl}} - 2}}$$

En un diseño de estas características (dos grupos aleatorios) también puede estimarse el tamaño del efecto mediante una cuantificación del grado de relación existente entre las dos variables involucradas en el diseño. Para este propósito suele utilizarse el coeficiente de correlación de Pearson, que aplicado a una variable dicotómica (la variable que define los dos grupos) y una variable cuantitativa (la variable en la cual se están comparando los grupos) recibe el nombre de coeficiente de correlación biserial-puntual (R_{bp}).

El coeficiente R_{bp} puede obtenerse fácilmente a partir del estadístico T de Student para muestras independientes:

$$R_{\text{bp}} = T / \sqrt{T^2 + n_{\text{tto}} + n_{\text{ctrl}} - 2}$$

Pero debe tenerse en cuenta que los estadísticos d y R_{bp} son formas equivalentes de cuantificar el mismo efecto. Por tanto, ambos estadísticos contienen la misma información. De hecho, uno es función del otro:

$$d = 2 R_{\text{bp}} / \sqrt{1 - R_{\text{bp}}^2}$$

Diseño pre-post con grupos aleatorios

La principal debilidad de un *diseño pre-post de grupo único* está en la ausencia de un grupo control: sin un grupo control no es posible controlar los posibles efectos asociados al paso del tiempo. La principal debilidad de un *diseño de grupos aleatorios* es que, al margen de la equivalencia entre grupos que se consiga alcanzar con la asignación aleatoria, la falta de seguimiento no permite saber si el efecto de la intervención se prolonga en el tiempo.

Un *diseño pre-post con grupos aleatorios* permite superar ambas debilidades. La debilidad del *diseño pre-post de grupo único* se supera incluyendo un grupo control; la debilidad del diseño de grupos aleatorios se supera realizando dos mediciones, una de ellas antes (pre) de llevar a cabo la intervención y otra de ellas después (post) (Feingold, 2009; Shadish, Cook, & Campbell, 2002, pp. 108-110).

Los datos obtenidos con un diseño de estas características suelen analizarse con un modelo de análisis de varianza (ANOVA); concretamente, con el modelo de *dos factores con medidas repetidas en uno de ellos* (Pardo & San Martín, 2015), o *modelo de dos factores con medidas parcialmente repetidas* (Ato & Vallejo, 2015).

Independientemente del nombre que se le dé al modelo que permite analizar los datos obtenidos con este diseño, se tienen dos factores o variables independientes y una variable dependiente. Las variables independientes son, ambas, variables categóricas: el grupo (factor inter-sujetos o completamente aleatorizado) y el momento (factor intra-sujetos o de medidas repetidas). La variable dependiente es una variable cuantitativa en la que se desea comparar los grupos para valorar si se ha producido algún cambio.

En este escenario, la valoración del cambio se lleva a cabo analizando el efecto de la *interacción* entre los factores *grupo* y *momento* (Kirk, 2013, pp. 30-75; Pardo, Garrido, Ruiz, & San Martín, 2007). El efecto de la interacción es justamente

el que permite comparar el cambio observado (diferencia pre-post) en el grupo experimental con el cambio observado en el grupo control. Y para cuantificar la intensidad de ese efecto se utilizan medidas de asociación como omega cuadrado, ω^2 , la cual puede estimarse fácilmente a partir del estadístico F asociado al efecto de la interacción (Kirk, 2013, pp. 134-138; Winer, Brown, & Michels, 1991, pp. 405-414).

$$\hat{\omega}^2 = \frac{gl_{\text{interacción}} (F_{\text{interacción}} - 1)}{gl_{\text{interacción}} (F_{\text{interacción}} - 1) + N}$$

En su momento (tercer estudio) veremos que hay otras formas de estimar el tamaño del efecto en este tipo de diseños.

La evaluación del cambio basada en los individuos

A los estadísticos que se describen en este apartado los vamos a agrupar bajo la denominación de *índices de cambio fiable (reliable change indices)* (Beaton, Bombardier, Katz, & Wright, 2001; de Vet et al., 2006). Estos índices comparten con las medidas del tamaño del efecto del apartado anterior el objetivo de intentar medir el cambio a partir de las respuestas de los sujetos obtenidas mediante escalas o cuestionarios. Pero las medidas del tamaño del efecto y los índices de cambio fiable abordan el estudio del cambio de formas muy distintas: mientras que las primeras ponen el foco en el cambio grupal (cambio en los centros de las distribuciones), los índices de cambio fiable ponen el foco en el cambio individual.

Esta diferencia entre ambos enfoques tiene importantes consecuencias. El hecho de que las medidas del tamaño del efecto se basen en la intensidad del efecto evaluado (tamaño de la diferencia entre promedios, intensidad de la relación entre variables) implica que únicamente es posible realizar una valoración del *cambio*

grupal. Los índices de cambio fiable, por el contrario, al poner el foco en el cambio individual, permiten conocer, en primer lugar, qué individuos concretos cambian (*cambio individual*) y, como consecuencia de esto, en segundo lugar, qué porcentaje de individuos cambian (*cambio grupal*).

Diseño pre-post de grupo único

Este diseño constituye el escenario en el que han surgido los estadísticos agrupados bajo la denominación de *índices de cambio fiable*. Algunos de ellos están basados en la estandarización de las diferencias pre-post, otros en el error típico de medida, otros en los pronósticos de la regresión lineal (Crosby et al., 2003; Ferrer & Pardo, 2014), etc. Pero todos ellos han sido diseñados para poder valorar si el cambio experimentado por un individuo concreto es lo suficientemente grande como para poder descartar que se deba a las variaciones aleatorias propias del azar muestral (Bauer, Lambert, & Nielsen, 2004; Crosby et al., 2003; Jacobson & Truax, 1991; Revicki, Hays, Cella, & Sloan, 2008).

En una exhaustiva revisión de los índices de cambio fiable, Pardo y Ferrer (2013) han identificado ocho de estos índices y han realizado una valoración de los mismos basada en la tasa de falsos positivos asociada a cada uno de ellos. En un trabajo posterior, Ferrer y Pardo (2014) han propuesto algunas modificaciones de los índices originales que han elevado a trece el número de índices disponibles. Tras valorar su comportamiento en diferentes condiciones, solamente algunos de ellos ofrecen un comportamiento aceptable.

Para nuestro trabajo, hemos decidido utilizar los dos índices de cambio fiable que mejores resultados han obtenido en la evaluación llevada a cabo por Ferrer y Pardo (2014):

- a. *Diferencia individual tipificada (standardized individual difference, SID)*. Este índice se obtiene dividiendo cada diferencia pre-post individual (D_i) entre las desviación típica de las diferencias pre-post (S_{dif}):

$$SID = \frac{D_i}{S_{dif}}$$

Por tanto, una *SID* no es más que una diferencia tipificada bajo el supuesto de que la media de las diferencias pre-post vale cero. Este estadístico fue propuesto inicialmente por Payne y Jones (1957) para evaluar el grado de discrepancia entre dos puntuaciones. Si la distribución de las diferencias pre-post es normal, el 95% de las *SID* se encontrarán entre los valores ± 1.96 , y el 90% entre ± 1.64 . Por lo tanto, cualquier cambio individual (diferencia pre-post) cuya *SID* alcance valores más extremos que estos se puede considerar como un *cambio fiable* (el umbral concreto depende del nivel de confianza que se elija y de si el contraste es bilateral o unilateral).

- b. *Índice de cambio fiable (Reliable Change Index, RCI)*. Este es probablemente el más popular de todos los índices de cambio. Fue propuesto originalmente por Jacobson y sus colaboradores (Jacobson, Follette, & Revenstorf, 1984; Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991). Se obtiene dividiendo cada diferencia individual pre-post entre el error típico de medida del instrumento.

De las diferentes versiones disponibles de este índice, aquí utilizaremos la que han propuesto Jacobson y Truax (1991) tras aplicar las correcciones sugeridas por Christensen y Mendoza (1986) y las recomendaciones de Maassen (2004). Por tanto usaremos una versión en la que no se asume igualdad entre las varianzas del pre y del post:

$$RCI = \frac{D_i}{\sqrt{(S_{pre} \sqrt{1 - R_{XX}})^2 + (S_{post} \sqrt{1 - R_{XX}})^2}}$$

donde R_{XX} se refiere a la fiabilidad del instrumento de medida. En la literatura estadística se han descrito diferentes versiones a partir de la formulación original de Jacobson y sus colaboradores. Aunque todas ellas parten de la misma idea, difieren en la manera de calcular el error típico de medida; más concretamente, difieren en la forma de calcular el coeficiente de fiabilidad. Algunos autores han propuesto estimar la fiabilidad mediante el coeficiente alfa de Cronbach (Martinovich, Saunders, & Howard, 1996; Tingey, Lambert, Burlingame, & Hansen, 1996). Sin embargo, Ferrer y Pardo (2014) han demostrado que estimar la fiabilidad a partir de la consistencia interna ofrece peores resultados (mayor tasa de falsos positivos) que estimarla a partir de la correlación entre las puntuaciones del pre y del post.

Para decidir qué valor de RCI indica que se ha producido un cambio fiable se han utilizado diferentes puntos de corte, pero los más habituales suelen ser ± 1.64 y ± 1.96 , que son los valores entre los que se encuentra un área del 90% y del 95%, respectivamente, en una curva normal tipificada.

Tanto SID como RCI son estadísticos *individuales*, es decir, estadísticos que ofrecen un valor concreto (una diferencia pre-post tipificada) para *cada individuo*. Los casos cuya puntuación SID o RCI supera un determinado punto de corte previamente establecido se considera que son casos que han experimentado un cambio fiable. Por tanto, tanto SID como RCI son estadísticos que informan sobre el *cambio individual*.

Pero, además, simplemente dividiendo el número de casos que experimentan un cambio fiable entre el número total de casos tratados, puede obtenerse el *porcentaje de cambios fiables*. Y este estadístico constituye un excelente indicador de *cambio grupal* que, además, es una forma alternativa de estimar el tamaño del efecto en un diseño *pre-post*.

Una de las principales ventajas de este último estadístico frente a otras medidas de cambio grupal es que el *porcentaje de cambios* puede interpretarse de forma sencilla sin tener grandes conocimientos de estadística o análisis de datos: el porcentaje de cambios tiene un mínimo y un máximo, y es más fácil de interpretar que un número expresado en escala de desviaciones típicas (como la diferencia tipificada d de Cohen): en general, decir que la diferencia entre los centros de las distribuciones vale 1.20 desviaciones típicas no es tan revelador como decir que se ha encontrado un 85% de recuperaciones.

Diseño de grupos aleatorios

Contrariamente a lo que pudiera parecer (pues ahora no se mide a los mismos sujetos dos veces), en un diseño de estas características también es posible dirigir la atención hacia el *cambio individual*. Delaney y Vargha (2002; 2000), por ejemplo, han propuesto el *estadístico A* basándose en la *medida del lenguaje común* (CL) descrita por McGraw y Wong (1992). Más tarde, estos estadísticos y otros similares han sido agrupados bajo la denominación general de *Probabilidad de superioridad* (*Probability of superiority, PS*) (Grissom & Kim, 2012, pp. 149-165).

Estos estadísticos se basan en la siguiente idea: si el grupo tratado no difiere del grupo control (es decir, si no existe efecto del tratamiento), la probabilidad de que una puntuación cualquiera del grupo tratado sea mayor que una puntuación cualquiera del grupo control [$P(Y_{\text{tto}} > Y_{\text{ctrl}})$] será la misma que la probabilidad de

que una puntuación cualquiera del grupo tratado sea menor que una puntuación cualquiera del grupo control [$P(Y_{\text{tto}} < Y_{\text{ctrl}})$]; es decir, $PS = 0.5$.

La forma lógica de estimar esta probabilidad consiste en tomar todas las posibles parejas de puntuaciones (una del grupo tratado y otra del grupo control) y en compararlas en la variable relevante (Y). El número de veces que la puntuación del grupo tratado superara a la de su par del grupo control permitiría obtener una estimación de PS . En este escenario, el valor de PS tenderá tanto más a uno cuanto mayores sean las puntuaciones del grupo tratado respecto de las del grupo control; y tenderá tanto más a cero cuanto menores sean las puntuaciones del grupo tratado respecto de las del grupo control.

No obstante, el valor de PS puede calcularse de distintas maneras dependiendo de los supuestos que se asuman sobre la variable Y . Dado que en este trabajo estudiaremos diferentes tipos de distribuciones (incluyendo algunas fuertemente asimétricas) y tamaños muestrales reducidos, optaremos por utilizar la versión de Delaney y Vargha (2002). Se trata de una versión no paramétrica basada, no en las puntuaciones originales, sino en sus correspondientes rangos:

$$PS = P(Y_{\text{tto}} > Y_{\text{ctrl}}) = [(\#(Y_{\text{tto}} > Y_{\text{ctrl}}) + 0.5 \#(Y_{\text{tto}} = Y_{\text{ctrl}}))] / (n_{\text{tto}} n_{\text{ctrl}})$$

Por tanto, PS se calcula contando el número de veces ($\#$) que una puntuación Y_{tto} es mayor que una puntuación Y_{ctrl} , ponderando los empates por la mitad y dividiendo ese resultado entre el número de comparaciones posibles.

Calculado de esta forma, el estadístico PS se corresponde exactamente con el área bajo la curva COR calculada con el método de Hanley y McNeil (1982). Consecuentemente, el cálculo de PS se puede simplificar utilizando las sumas de rangos que se emplean, por ejemplo, en el cálculo del estadístico U de Mann-Whitney:

$$PS = [n_{tto} n_{ctrl} + n_j (n_j + 1) / 2 - SUM_j] / (n_{tto} n_{ctrl})$$

donde *SUM* se refiere a la suma de los rangos y *j* se refiere al grupo al que corresponde la suma de rangos menor.

Aunque para calcular *PS* es necesario contar con información relativa a todo el grupo (en concreto, es necesario conocer el tamaño de los grupos), su cálculo se basa en la comparación de cada puntuación individual del grupo tratado con cada puntuación individual del grupo control. Por tanto, se trata de un estadístico que informa sobre el *cambio individual*, o más concretamente, sobre la probabilidad de que una puntuación aleatoriamente seleccionada del grupo tratado se encuentre por encima de una puntuación aleatoriamente seleccionada del grupo control.

Una de las principales ventajas de este estadístico frente a otros más habituales en este tipo de diseños (la *d* de Cohen, por ejemplo) es que, al igual que el *porcentaje de cambios fiables* que se obtiene en un diseño pre-post de grupo único, puede ser interpretado de forma sencilla sin tener grandes conocimientos de estadística o análisis de datos: al expresar el efecto de una intervención como una probabilidad o como un porcentaje se está trabajando en una escala que tiene valores mínimo y máximo, y sobre todo en una métrica que, desde el punto de vista aplicado, resulta más intuitivo para los profesionales aplicados que se ven en la necesidad de valorar la intensidad del efecto de una intervención.

Diseño pre-post con grupos aleatorios

Exactamente igual que ocurriría con la valoración del *cambio grupal* en este diseño, la valoración del *cambio individual* requiere comparar lo que ocurre en el grupo tratado con lo que ocurre en el grupo control. Pero ahora, el hecho de contar con dos mediciones facilita bastante las cosas.

La diferencia pre-post de cada individuo ya está informando de la posible presencia de un cambio individual. No obstante, la presencia de un grupo control añade información extra que no puede pasarse por alto: los cambios del grupo control indican qué ocurre cuando no hay intervención. En consecuencia, la valoración del cambio exige que el eventual cambio en los sujetos no tratados sea descontado del cambio observado en el grupo tratado.

Lo que procede, por tanto, es obtener el *porcentaje neto de mejorías* (P_{NM}). Este porcentaje neto se obtiene restando los porcentajes netos de mejorías del grupo tratado y del grupo control. Y el porcentaje neto de cada grupo se obtiene restando el porcentaje de casos que empeoran (P^-) al porcentaje de casos que mejoran (P^+). Se llega así a la siguiente ecuación:

$$P_{NM} = (P_{tto}^+ - P_{tto}^-) - (P_{ctrl}^+ - P_{ctrl}^-)$$

Objetivos

Ya hemos señalado que el enfoque basado en los centros y el enfoque basado en los individuos valoran el cambio de forma muy distinta. Sin embargo, los resultados de Norman *et al.* (2001) sugieren que, al menos en un **diseño pre-post de grupo único** y bajo ciertas condiciones, es posible obtener parte de la información que ofrece un enfoque a partir de la información que ofrece el otro. Ahora bien, Norman y sus colaboradores no incluyen en su trabajo ninguna función que relacione el porcentaje de cambios individuales con el cambio en los centros de las distribuciones, ni informan del grado de ajuste de dicha función (lo cual sería muy útil para evaluar la calidad de sus estimaciones). En otras palabras, el único trabajo que conocemos sobre la relación entre el tamaño del efecto y el porcentaje de

cambios deja dos cuestiones importantes sin aclarar: en primer lugar, la función matemática que permite pronosticar el porcentaje de cambios a partir del cambio grupal promedio; y, en segundo lugar, la calidad de esa función. Además, Norman y sus colaboradores no ofrecen información sobre el proceso de simulación implementado en su estudio: no informan de cómo han simulado sus datos ni tampoco de las propiedades de las distribuciones resultantes de la simulación. En consecuencia, falta información relevante para poder evaluar la calidad de los datos sobre los que han basado sus conclusiones. Con todo, su trabajo constituye uno de los pocos intentos encontrados en la literatura estadística (quizá el único) destinados a relacionar la información sobre el cambio basado en los centros con la información sobre el cambio basado en los individuos. En el caso de un **diseño de grupos aleatorios** se ha descrito la relación entre la *diferencia estandarizada de Cohen* y la *probabilidad de superioridad*, pero solo en el caso de variables con distribución normal (Grissom, 1994; Grissom & Kim, 2012, p. 168). Y en lo relativo al **diseño pre-post con grupos aleatorios** no hemos encontrado ningún trabajo que estudie la relación entre el cambio evaluado a partir de los centros y el cambio evaluado a partir de los individuos.

Varios autores han señalado la importancia de poder intercambiar resultados entre el nivel grupal y el individual sin perder información en el camino (Beaton et al., 2001). Pero la escasez de estudios sobre el tema y la falta de conclusiones sólidas nos llevan a pensar que es necesario seguir investigando sobre la relación entre el cambio basado en los centros y el cambio basado en los individuos.

Con esto en mente, este trabajo tiene como **objetivo principal** explorar y describir la relación entre dos formas distintas de evaluar el cambio. Ya nos hemos referido a ellas como el enfoque basado en los centros y el enfoque basado en los individuos. Más concretamente nos proponemos:

1. Explorar y describir las eventuales equivalencias numéricas existentes entre las estimaciones de cambio realizadas con el enfoque basado en los centros y el enfoque basado en los individuos. Para ello, pretendemos encontrar la función o funciones capaces de describir la relación entre la información que ofrecen ambos enfoques, y estudiar su forma y su ajuste. Todo esto, en diseños pre-post de grupo único, en diseños de grupos aleatorios y en diseños pre-post con grupos aleatorios. Este objetivo incluye estudiar bajo qué condiciones es posible afirmar que la información que ofrecen ambos enfoques está relacionada.
2. Con apoyo de las funciones referidas en el punto anterior, establecer criterios objetivos para evaluar una intervención basados en cada uno de los métodos, y describir las equivalencias entre estos criterios. Esto implica valorar la adecuación de algunos criterios frecuentemente utilizados en la literatura estadística para evaluar una intervención. En su trabajo de 1988, Cohen propuso de forma tentativa algunos referentes para considerar un efecto como pequeño, mediano o grande. Por ejemplo, en el caso del estadístico d , consideró que efectos pequeños se corresponderían con valores en torno a 0.2, efectos medios con valores en torno a 0.5 y efectos grandes con valores en torno a 0.8 (Cohen, 1969, pp. 24-27, 1992). Aunque el propio Cohen ha señalado la conveniencia de interpretar cada efecto en su contexto, comparándolo con otros efectos encontrados en estudios previos, en la práctica sus puntos de corte han sido y son utilizados con gran frecuencia para valorar todo tipo de efectos en intervenciones aplicadas. En este trabajo estudiaremos la correspondencia de estos criterios, basados en el cambio en los centros, con los criterios basados en el cambio individual, y valoraremos su adecuación y pertinencia.

Estudio 1

Relación entre el cambio basado en los centros y el cambio basado en los individuos.

I. El caso de los diseños pre-post de grupo único

Objetivo

El objetivo general de este estudio es explorar la relación entre diferentes métodos de evaluación del cambio. Más concretamente, la relación entre la información que se obtiene con el enfoque basado en los centros de las distribuciones y la que se obtiene con el enfoque basado en las puntuaciones individuales.

Este propósito general de exploración de la relación entre ambos enfoques podemos concretarlo en tres objetivos más específicos: (a) averiguar si existe relación entre el tamaño del efecto (obtenido a partir del cambio en los centros de las distribuciones) y el porcentaje de cambios (obtenido a partir de la acumulación de cambios individuales); (b) encontrar la forma de esa relación, la función matemática que la representa y su grado de ajuste; y (c) determinar bajo qué condiciones (distribución de las puntuaciones, correlación pre-post y tamaño muestral) es posible afirmar que tal relación existe.

En este primer estudio comenzaremos explorando la relación entre ambos enfoques en el marco de los diseños pre-post de grupo único. Es bien sabido que, para evaluar adecuadamente la eficacia de un tratamiento, es necesario incorporar al diseño un grupo control (Shadish et al., 2002, pp. 108-110). Sin embargo, hemos seleccionado un diseño pre-post de grupo único como punto de partida ya que este es el escenario en el que se han desarrollado todos los índices utilizados para evaluar el cambio individual (Crawford et al., 1998; Hageman & Arrindell, 1999; Jacobson & Truax, 1991; Payne & Jones, 1957; Wyrwich et al., 1999).

Método

Para estudiar la relación entre el tamaño del efecto y el porcentaje de cambios hemos simulado datos correspondientes a un diseño pre-post de grupo único (es decir, sin grupo control). Por tanto, hemos adoptado un escenario en el que la misma variable se mide en los mismos sujetos en dos momentos distintos (generalmente antes y después de la intervención) con el objetivo de evaluar si se ha producido un cambio entre las dos mediciones.

Para definir las diferentes condiciones de la simulación, hemos manipulado cuatro criterios:

- a. *Tamaño del efecto* (δ). El efecto de la intervención se ha controlado modificando el valor de la diferencia estandarizada δ propuesta por Cohen (1969):

$$\delta = \mu_{\text{dif}} / \sigma_{\text{dif}}$$

donde μ_{dif} es la media de las diferencias individuales post-pre y σ_{dif} es la desviación típica esas dichas diferencias. Como ya hemos señalado

anteriormente, en este trabajo hemos optado por utilizar σ_{dif} como estandarizador por considerarlo mejor para nuestros objetivos que la desviación típica de las puntuaciones pre, propuesta por algunos autores (Cohen, 1969; Cumming, 2012; Cumming & Finch, 2001).

Hemos elegido efectos de 13 tamaños distintos: de 0 a 3.6 puntos, a incrementos de 0.3 puntos (en la métrica de las diferencias pre-post). Un efecto de, por ejemplo, tamaño 0.6 indica que la media de las diferencias pre-post es 0.6 veces la desviación típica de las diferencias pre-post. Hemos elegido este rango tan amplio de efectos, desde un efecto nulo a uno extremadamente grande, para permitir que el porcentaje de cambios individuales pudiera abarcar todo su rango (0% - 100%).

En nuestros cálculos hemos asumido que el efecto de la intervención hace aumentar la media del momento post. En consecuencia, hemos trabajado con efectos positivos (se espera que la resta de las medias post menos pre lleve a un resultado mayor de cero) y contrastes unilaterales derechos. En relación con esto, conviene advertir que, dado que los contrastes bilaterales son preferibles cuando no se conoce *a priori* la dirección del efecto de la intervención y teniendo en cuenta que en nuestra simulación generaremos solo efectos positivos, emplearemos contrastes bilaterales únicamente en la condición en la que se impone un efecto nulo ($\delta = 0$).

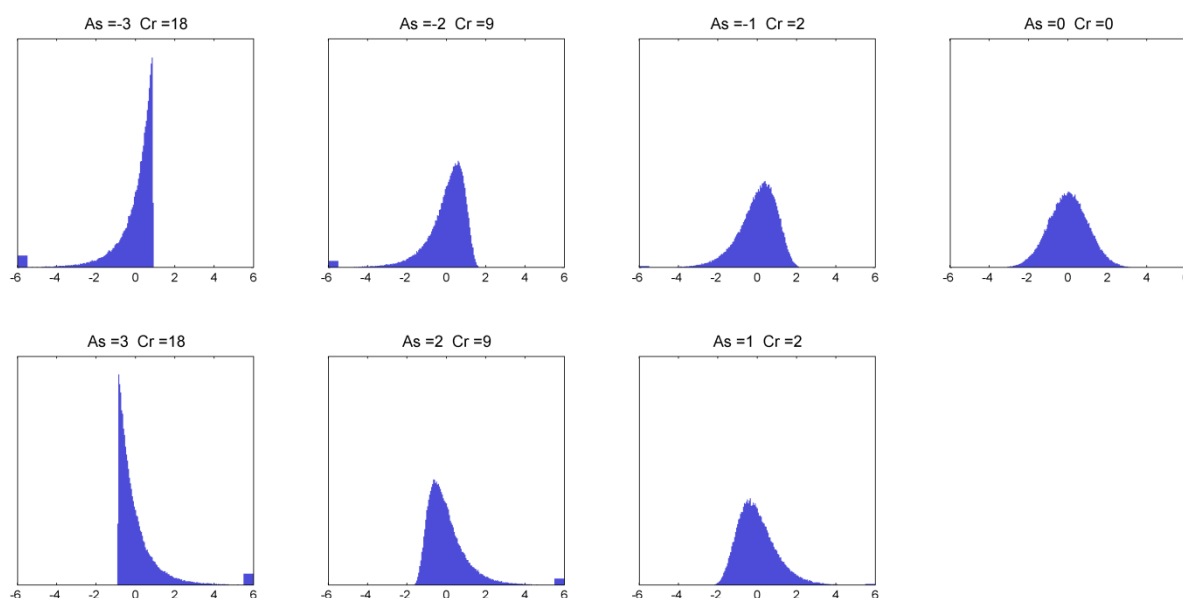
- b. Tamaño muestral (n).* Hemos seleccionado tres tamaños muestrales (25, 50 y 100) para simular lo que en el ámbito clínico suele considerarse muestras pequeñas medianas y grandes (Crawford & Howell, 1998).

- c. *Correlación pre-post* (ρ_{XY}): 0.5, 0.7, 0.9. Hemos elegido estos valores para simular fiabilidades o correlaciones pre-post moderadas, medias y altas (Nunnally & Bernstein, 1994, pp. 264-265; Pedhazur & Schmelkin, 1991, pp. 109-110). Para cuantificar el grado de relación hemos utilizado el coeficiente de correlación de Pearson.
- d. *Forma de las distribuciones*. Dado que en contextos aplicados es bastante habitual encontrar distribuciones que se desvían de la normalidad de forma moderada e incluso severa (Blanca et al., 2013; Micceri, 1989), hemos decidido simular distribuciones con diferente grado de alejamiento de la normalidad. Para esto, hemos controlado el grado de asimetría (γ_1) y el grado de curtosis (γ_2). Más concretamente, hemos simulado siete distribuciones que van desde asimetría negativa extrema ($\gamma_1 = -3$) hasta asimetría positiva extrema ($\gamma_1 = 3$):
1. Asimetría negativa extrema: $\gamma_1 = -3$, $\gamma_2 = 18$.
 2. Asimetría negativa moderada: $\gamma_1 = -2$, $\gamma_2 = 9$.
 3. Asimetría negativa leve: $\gamma_1 = -1$, $\gamma_2 = 2$.
 4. Normalidad: $\gamma_1 = 0$, $\gamma_2 = 0$.
 5. Asimetría positiva leve: $\gamma_1 = 1$, $\gamma_2 = 2$.
 6. Asimetría positiva moderada: $\gamma_1 = 2$, $\gamma_2 = 9$.
 7. Asimetría positiva extrema: $\gamma_1 = 3$, $\gamma_2 = 18$.

Para interpretar correctamente el grado de curtosis elegido en cada condición es necesario tener en cuenta que el grado de curtosis de una distribución está parcialmente condicionado por su grado de asimetría. La Figura 1 muestra las siete formas de la distribución que genera cada combinación de γ_1 y γ_2 .

Figura 1.

Formas de la distribución generadas en las siete condiciones simuladas. Los casos alejados del centro en más 5,5 desviaciones típicas se han agrupado en el extremo del correspondiente histograma.



Combinando estos cuatro criterios hemos creado $13 \times 3 \times 3 \times 7 = 819$ condiciones de simulación diferentes. En cada una de estas condiciones se han generado 500 muestras (409500 muestras en total). Para realizar esta simulación hemos utilizado el programa MatLab 2011a. El procedimiento concreto se explica en el próximo párrafo (y el código de MatLab correspondiente se puede consultar en el Apéndice 2).

En primer lugar, hemos generado una matriz $\mathbf{X} = (X^*, Y^*)$ con n parejas de puntuaciones en dos variables independientes entre sí ($X = \text{pre}$; $Y = \text{post}$). Las puntuaciones se han generado mediante el sistema de distribuciones de Pearson (función *pearsrnd* en MatLab). Ambas variables se han creado con la misma media, desviación típica, asimetría y curtosis. La media se ha fijado en cero y la desviación típica en uno en todas las condiciones simuladas. La asimetría y la curtosis se han

modificado sistemáticamente aplicando los valores de γ_1 y γ_2 detallados anteriormente. El hecho de que X e Y se hayan generado aleatoriamente garantiza que, tal como ocurre en contextos reales, la puntuación del momento post (Y) pueda ser igual, mayor, mucho mayor, menor o mucho menor que su correspondiente puntuación en el momento pre.

En segundo lugar, hemos impuesto el grado de relación entre las variables de la matriz \mathbf{X} mediante la descomposición de covarianza de Cholesky (función *cholcov* en MatLab) de la matriz de correlaciones \mathbf{R} correspondiente a la correlación elegida (ρ_{XY}). La matriz resultante $\mathbf{M} = (X, Y)$ contenía dos variables con valores de asimetría, curtosis, y correlación R_{XY} similares a los especificados para cada condición. Con esta transformación se consigue asegurar que, tal como ocurre en contextos reales, las puntuaciones de los momentos pre y post no sean completamente independientes.

En el último paso del proceso de simulación se ha modificado la variable Y para hacer que su media se aproximase a la media elegida para cada condición simulada. Para esto, a cada valor individual de Y se le ha sumado, en cada muestra, el resultado de multiplicar la desviación típica de las diferencias pre-post de esa muestra por el tamaño del efecto (valor de δ) de la correspondiente condición.

Debe tenerse en cuenta que, aunque trabajar directamente con la distribución de las diferencias habría sido mucho más simple que simular las distribuciones de las puntuaciones pre y post, nuestro propósito era emular escenarios reales en los que la distribución de las diferencias raramente se conoce.

También debe tenerse en cuenta que el proceso de simulación utilizado genera puntuaciones pre y post homocedásticas. Aunque en contextos aplicados no hay razón para asumir que las varianzas serán iguales antes y después de una

intervención, decidimos asumir igualdad entre las varianzas del pre y del post para comenzar con el escenario más simple.

Análisis de los datos

En primer lugar, tras generar las 409500 muestras, hemos comprobado si las propiedades de las distribuciones simuladas eran congruentes con los valores especificados para generarlas. Para ello, hemos calculado, en cada condición simulada, estadísticos descriptivos de tendencia central, de dispersión, de asimetría, de curtosis y la correlación pre-post.

En segundo lugar, hemos calculado, en cada muestra, el tamaño del efecto empírico dividiendo la diferencia de las medias post-pre entre la desviación típica de las diferencias. Este tamaño del efecto empírico es el que se ha utilizado para cuantificar el cambio grupal en cada muestra.

En tercer lugar, hemos calculado los dos índices de cambio individual elegidos para cuantificar el cambio individual en diseños *pre-post* de grupo único. Tal como hemos señalado en la introducción, estos dos índices son los que servirán para identificar qué puntuaciones individuales experimentan un cambio fiable:

a. Diferencia Individual tipificada (standardized individual difference):

$$SID = D_i / S_{\text{dif}}$$

donde D_i es la diferencia post-pre correspondiente a cada individuo y S_{dif} es la desviación típica de las n diferencias individuales de cada muestra.

b. *Índice de cambio fiable (reliable change index),*

$$RCI = \frac{D_i}{\sqrt{(S_{\text{pre}} \sqrt{1 - R_{XX}})^2 + (S_{\text{post}} \sqrt{1 - R_{XX}})^2}}$$

donde R_{XX} se refiere a la fiabilidad del instrumento de medida, la cual, según hemos explicado ya, no estimaremos a partir de la consistencia interna (alfa de Cronbach) sino a partir de la correlación test-retest (es decir, mediante el coeficiente de correlación de Pearson entre las puntuaciones del pre y del post). Ahora que ya sabemos que, al simular los datos, hemos llamado X al momento pre e Y al momento post, en la ecuación correspondiente a RCI puede utilizarse S_X en lugar de S_{pre} y S_Y en lugar de S_{post} .

Estos dos estadísticos se aplican a cada par de puntuaciones de cada muestra, es decir, a cada una de las diferencias post-pre de cada muestra; por tanto, estos estadísticos generan un valor concreto para cada individuo en cada muestra. Si la distribución de las diferencias pre-post es normal, el 95% de los valores SID y RCI se encontrarán dentro del rango ± 1.96 , y el 90% dentro del rango ± 1.645 . Por lo tanto, en las condiciones en las que hemos impuesto un efecto positivo, hemos considerado que se había producido un cambio fiable cuando SID o RCI han superado el punto de corte 1.645 (estrategia equivalente a un contraste unilateral derecho con un nivel de confianza del 95%); y en las condiciones en las que se ha impuesto un efecto nulo, hemos considerado que se había producido un cambio fiable cuando SID o RCI han tomado un valor fuera del rango ± 1.96 (estrategia equivalente a un contraste bilateral con un nivel de confianza del 95%).

Basándonos en la información individual ofrecida por SID y RCI hemos calculado el porcentaje de cambios individuales *de cada muestra* simplemente

dividiendo el número valores *SID* o *RCI* por encima del punto de corte entre el correspondiente tamaño muestral.

Por último, hemos utilizado cada *tamaño de efecto empírico* y cada *porcentaje de cambios individuales* (500 pares de valores para cada condición simulada) para obtener: (a) un diagrama de dispersión para inspeccionar el tipo de relación subyacente entre los dos estadísticos; y (b) tres funciones diferentes de regresión (lineal, cuadrática y cúbica) para cuantificar el grado en el que el cambio en el centro de la distribución permite predecir el porcentaje de cambios individuales. Estas funciones fueron estimadas mediante la sentencia *curvefit* del paquete SPSS Statistics v20.

Resultados

Dado que se han generado 819 condiciones distintas, incluir los resultados de todas ellas en esta sección sería poco práctico. Por esta razón aquí únicamente se incluye una parte de los resultados. No obstante, los resultados elegidos sirven como resumen y representación del conjunto completo. El resto de resultados pueden consultarse en el siguiente enlace:

<https://copy.com/mPkovNoNQyHAXwCG>

En la Tabla 1 se recoge un resumen descriptivo de los datos simulados en cada una de las siete condiciones con $n = 25$, $\rho_{XY} = 0.7$ y $\delta = 0.9$. Se incluyen las medias (M), desviaciones típicas (S) de las puntuaciones del pre (X), del post (Y) y de las diferencias post-pre (dif). También se incluyen los índices de asimetría y curtosis, el coeficiente de correlación de Pearson entre las puntuaciones pre y post (R_{XY}) y el tamaño del efecto empírico (d). Los resultados del resto de condiciones de la simulación son similares a estos, y no proporcionan información adicional de interés.

Tabla 1.

Estadísticos descriptivos correspondientes a las siete distribuciones incluidas en el estudio (con $n = 25$, $\rho_{XY} = 0.7$, $\delta = 0.9$). Cada casilla de la tabla incluye la media (desv. típica) de las 500 muestras simuladas en cada condición

	$As = -3$ $Cr = 18$	$As = -2$ $Cr = 9$	$As = -1$ $Cr = 2$	$As = 0$ $Cr = 0$	$As = 1$ $Cr = 2$	$As = 2$ $Cr = 9$	$As = 3$ $Cr = 18$
M_X	.00 (.20)	-.01 (.20)	-.02 (.20)	.02 (.21)	.02 (.21)	-.02 (.19)	-.01 (.20)
M_Y	.67 (.17)	.67 (.18)	.68 (.19)	.71 (.23)	.69 (.28)	.63 (.32)	.68 (.38)
M_{dif}	.66 (.16)	.68 (.16)	.70 (.16)	.69 (.19)	.67 (.23)	.65 (.26)	.69 (.32)
S_X	.94 (.31)	.98 (.28)	.99 (.19)	1.00 (.15)	.99 (.19)	.93 (.26)	.92 (.33)
S_Y	.96 (.26)	.98 (.23)	1.00 (.18)	.99 (.14)	.99 (.17)	.96 (.23)	.96 (.27)
S_{dif}	.74 (.22)	.75 (.18)	.76 (.14)	.77 (.11)	.75 (.13)	.73 (.18)	.75 (.24)
$As\ X$	-1.76 (.75)	-1.27 (.81)	-.76 (.59)	-.01 (.49)	.74 (.63)	1.14 (.71)	1.68 (.76)
$As\ Y$	-1.32 (.65)	-.97 (.69)	-.55 (.53)	.02 (.45)	.53 (.55)	.93 (.68)	1.37 (.72)
$As\ dif$	-1.20 (.97)	-.85 (.83)	-.53 (.64)	.01 (.47)	.48 (.60)	.86 (.82)	1.28 (.98)
$Cr\ X$	3.79 (3.94)	2.42 (3.82)	.88 (2.00)	-.01 (.90)	.87 (2.16)	1.76 (3.10)	3.37 (4.07)
$Cr\ Y$	2.09 (2.93)	1.48 (2.77)	.45 (1.50)	-.02 (.84)	.40 (1.71)	1.41 (2.61)	2.43 (3.45)
$Cr\ dif$	3.25 (3.59)	1.85 (2.79)	.74 (1.90)	.02 (.84)	.60 (1.58)	1.75 (2.98)	3.40 (3.74)
R_{XY}	.68 (.17)	.69 (.16)	.70 (.12)	.69 (.11)	.70 (.13)	.69 (.14)	.66 (.18)
d	.93 (.22)	.93 (.21)	.93 (.21)	.89 (.21)	.88 (.21)	.87 (.21)	.88 (.21)

A medida que el tamaño muestral aumenta, los datos simulados se acercan más a los valores de referencia utilizados para generarlos. Por lo tanto, los resultados obtenidos con $n = 25$ permiten formarse una idea bastante acertada de la calidad de la simulación llevada a cabo.

Los resultados de la Tabla 1 permiten apreciar que las propiedades de las muestras generadas se corresponden con las propiedades impuestas en la simulación. Así, el tamaño del efecto d vale aproximadamente 0.9; la media del momento pre (M_X) toma un valor muy próximo a cero; la media del momento post (M_Y) toma un valor muy próximo a 0.7 (es decir, al valor de μ_X más δ veces la desviación típica de las diferencias); las desviaciones típicas de los momentos pre y

post (S_X y S_Y) son aproximadamente iguales a uno; el coeficiente de correlación de Pearson (R_{XY}) toma un valor muy próximo a 0.7; y la desviación típica de las diferencias (S_{dif}) toma un valor muy próximo a

$$[\sigma^2_X + \sigma^2_Y - 2\sigma_{XY}]^{1/2} = [1 + 1 - 2(.7)]^{1/2} = 0.775$$

Únicamente los índices de asimetría y de curtosis de las distribuciones simuladas difieren ligeramente de los valores impuestos. Pero esto es algo esperable ya que, con tamaños muestrales pequeños, es altamente improbable obtener valores extremos de asimetría y curtosis, incluso aunque las poblaciones muestreadas sean extremas en ambas propiedades.

En la Tabla 2 se muestran los resultados relativos a uno de los dos índices de cambio utilizados: la diferencia individual tipificada (*SID*). Los valores de esta tabla se han obtenido promediando primero los n valores *SID* de cada muestra (debe tenerse en cuenta que a cada caso simulado le corresponde un valor *SID*) y calculando a continuación la media y desviación típica de cada una de esas 500 medias. En la Tabla 2 se está informando únicamente de las 91 condiciones correspondientes a $n = 25$ y $\rho_{XY} = 0.7$.

Los resultados relativos al estadístico *RCI* son prácticamente idénticos a los presentados para *SID*. Y los resultados relativos al resto de condiciones simuladas son muy similares a los de la Tabla 2 (no ofrecen información relevante adicional). Sin embargo, toda la información relativa a *RCI* y al resto de condiciones simuladas puede consultarse en <https://copy.com/mPkovNoNQyHAXwCG>

Dado que el estadístico *SID* se calcula dividiendo cada diferencia post-pre entre la desviación típica de las diferencias, es esperable que la media de los n valores *SID* de cada muestra coincida con el tamaño del efecto asignado a la correspondiente condición. Los resultados de la Tabla 2 indican que eso es justamente lo que ocurre; y, curiosamente, ocurre independientemente de la forma

de la distribución de las puntuaciones. Además, conforme aumenta el tamaño muestral, esta correspondencia también aumenta.

Tabla 2.

Medias (desviaciones típicas) del estadístico *SID* en todas las distribuciones simuladas con $n = 25$ y $\rho_{XY} = 0.7$

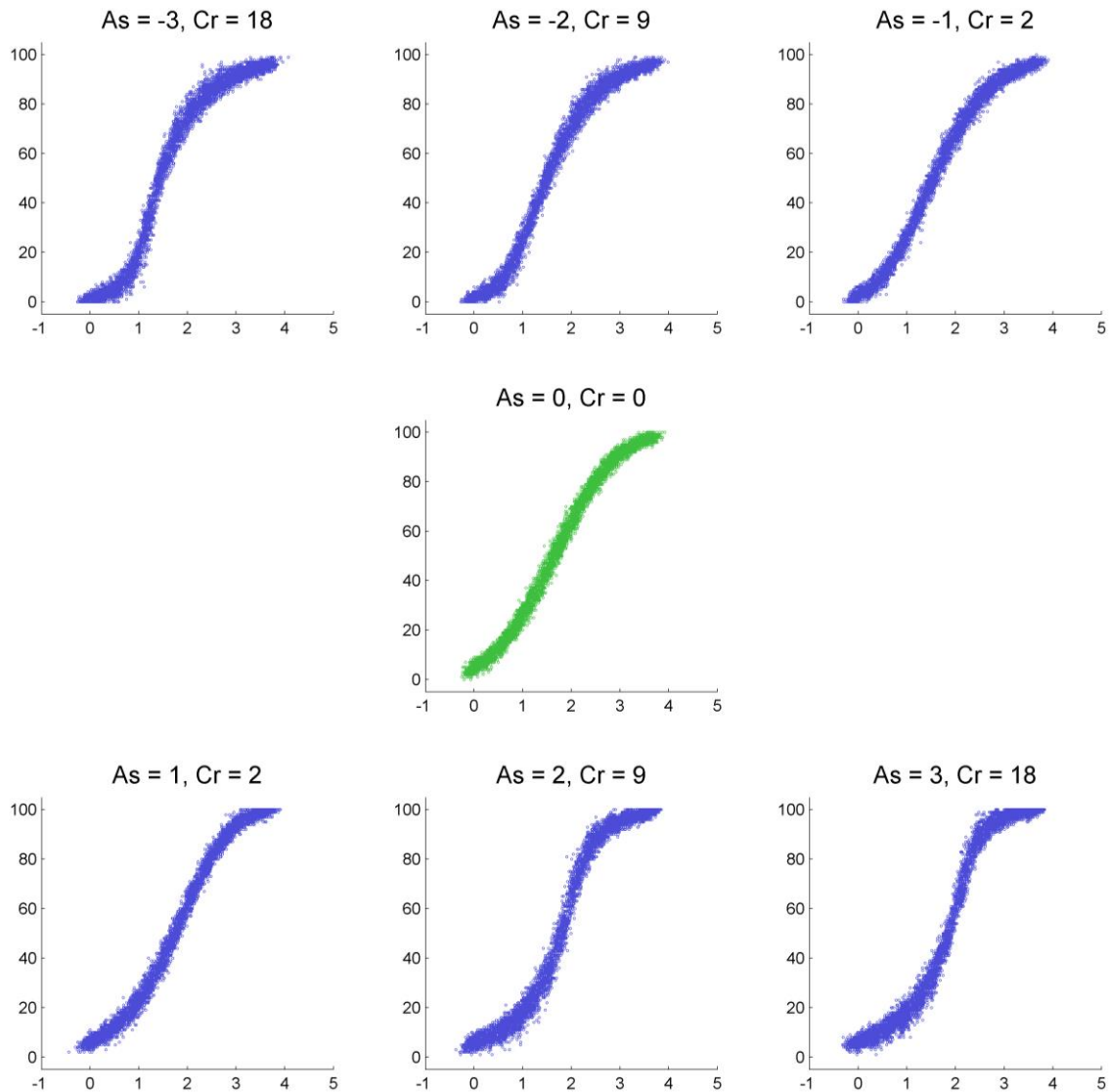
δ	<i>Forma de la distribución</i>						
	<i>As = -3 Cr = 18</i>	<i>As = -2 Cr = 9</i>	<i>As = -1 Cr = 2</i>	<i>As = 0 Cr = 0</i>	<i>As = 1 Cr = 2</i>	<i>As = 2 Cr = 9</i>	<i>As = 3 Cr = 18</i>
.0	.05 (.24)	.04 (.22)	.02 (.20)	-.01 (.20)	.00 (.22)	-.03 (.21)	-.04 (.23)
.3	.34 (.23)	.33 (.22)	.31 (.21)	.30 (.21)	.29 (.20)	.30 (.22)	.27 (.22)
.6	.65 (.23)	.63 (.23)	.61 (.20)	.59 (.21)	.59 (.21)	.57 (.22)	.58 (.22)
.9	.93 (.22)	.93 (.21)	.93 (.21)	.89 (.21)	.88 (.21)	.87 (.21)	.88 (.21)
1.2	1.24 (.22)	1.22 (.22)	1.21 (.21)	1.20 (.22)	1.20 (.21)	1.18 (.20)	1.17 (.23)
1.5	1.53 (.21)	1.54 (.21)	1.51 (.20)	1.49 (.20)	1.50 (.22)	1.47 (.21)	1.47 (.22)
1.8	1.84 (.23)	1.83 (.22)	1.81 (.21)	1.80 (.21)	1.77 (.20)	1.78 (.21)	1.76 (.22)
2.1	2.14 (.23)	2.13 (.21)	2.11 (.20)	2.11 (.22)	2.09 (.21)	2.07 (.22)	2.05 (.22)
2.4	2.44 (.22)	2.43 (.21)	2.41 (.21)	2.41 (.19)	2.38 (.21)	2.38 (.22)	2.36 (.23)
2.7	2.74 (.22)	2.72 (.22)	2.73 (.21)	2.70 (.22)	2.68 (.20)	2.68 (.22)	2.66 (.23)
3.0	3.03 (.23)	3.00 (.21)	3.03 (.22)	3.01 (.19)	2.98 (.20)	2.98 (.22)	2.97 (.23)
3.3	3.35 (.22)	3.32 (.21)	3.32 (.20)	3.31 (.20)	3.28 (.21)	3.27 (.22)	3.26 (.23)
3.6	3.64 (.22)	3.63 (.22)	3.60 (.22)	3.59 (.21)	3.60 (.21)	3.58 (.21)	3.55 (.23)

Tras estudiar la correspondencia existente entre los datos simulados y los valores impuestos en la simulación, hemos pasado a estudiar la relación entre el tamaño del efecto (cambio en el centro de la distribución de las diferencias) y el porcentaje de cambios (acumulación de cambios individuales). Para ello, hemos comenzado construyendo varios diagramas de dispersión con el *tamaño del efecto* en el eje horizontal y el *porcentaje de cambios* en el eje vertical. En la Figura 2 se muestran algunos de estos diagramas de dispersión: los basados en *SID*, con $n = 100$ y $\rho_{XY} = 0.7$. En estos diagramas, cada punto corresponde a una de las muestras

simuladas: los 13 tamaños de efecto multiplicados por las 500 muestras simuladas dan lugar a 6500 puntos por diagrama. Para facilitar la interpretación de la relación subyacente, hemos decidido mostrara aquí los diagramas correspondientes a las condiciones con mayor tamaño muestral. No obstante, con $n = 25$ y $n = 50$ se encuentran exactamente los mismos patrones, aunque con una variabilidad mayor: para un determinado efecto, muestras más pequeñas conllevan mayor variabilidad en el porcentaje de cambios, manteniéndose constante el valor central. Con $\rho_{XY} = 0.5$ y $\rho_{XY} = 0.9$ se encuentran los mismos patrones.

Figura 2.

Diagramas de dispersión con el tamaño del efecto (d) en el eje horizontal y el porcentaje de cambios (basado en *SID*) en el eje vertical ($n = 100$, $\rho_{XY} = 0.7$).



Los resultados basados en el estadístico *RCI* son muy similares a los encontrados con el estadístico *SID*. Se observa el mismo patrón en la relación, con un efecto adicional: a medida que aumenta la asimetría (positiva o negativa), el porcentaje de cambios de algunas muestras alcanza un valor por encima de la tendencia

general. En otras palabras, cuando la asimetría es extrema, algunas muestras muestran un porcentaje de cambios mayor que el resto.

Para poder cuantificar la relación que se aprecia en los diagramas de la Figura 2, estimamos tres modelos de regresión diferentes: lineal, cuadrático y cúbico. Nótese que los diagramas de la Figura 2 sugieren que la relación subyacente sería adecuadamente descrita utilizando una función logística o probit. Sin embargo, el ajuste que se obtiene con estas dos funciones es peor que el que se obtiene con las funciones lineal, cuadrática o cúbica. Además, para poder estimar las funciones logística y probit, sería necesario dicotomizar la variable dependiente (el porcentaje de cambios), y al hacer esto los valores pronosticados ya no se referirían al porcentaje de cambios, sino a la probabilidad de pertenecer a una de las dos categorías de la variable dicotomizada.

En todos los casos se ha utilizado el tamaño del efecto como variable independiente y el porcentaje de cambios como variable dependiente. Las tres funciones fueron estimadas para cada una de las 63 condiciones resultantes de la combinación de los tres tamaños muestrales, las tres fiabilidades pre-post, y las siete condiciones de asimetría.

En la Tabla 3 se informa de los coeficientes de determinación (R^2) para las tres funciones. Los resultados de la tabla están basados en los porcentajes de cambio obtenidos con el estadístico *SID* y con $n = 25$. Puesto que la dispersión en los diagramas disminuye al aumentar el tamaño muestral, los valores de R^2 recogidos en la Tabla 3 son los peores (los más bajos) de todos los obtenidos; es decir, con $n = 50$ y $n = 100$ el ajuste es siempre mejor. Con todo, incluso con $n = 25$, las tres funciones permiten alcanzar un ajuste excelente.

Tabla 3.

Valores de R^2 obtenidos con las funciones lineal, cuadrática y cúbica ($n = 25$, porcentajes de cambio basados en el estadístico *SID*)

Distribución	ρ_{XY}	Lineal	Cuadrática	Cúbica
$As = -3, Cr = 18$.5	.909	.939	.961
	.7	.908	.940	.960
	.9	.899	.940	.957
$As = -2, Cr = 9$.5	.933	.954	.972
	.7	.931	.954	.971
	.9	.927	.954	.970
$As = -1, Cr = 2$.5	.951	.963	.979
	.7	.949	.961	.978
	.9	.947	.962	.978
$As = 0, Cr = 0$.5	.962	.965	.983
	.7	.962	.965	.982
	.9	.962	.964	.982
$As = 1, Cr = 2$.5	.956	.956	.980
	.7	.955	.955	.980
	.9	.954	.955	.979
$As = 2, Cr = 9$.5	.944	.944	.975
	.7	.943	.944	.976
	.9	.940	.942	.974
$As = 3, Cr = 18$.5	.924	.926	.968
	.7	.921	.925	.964
	.9	.915	.921	.962

En primer lugar, la función lineal toma valores de R^2 alrededor de 0.90 en distribuciones con asimetría negativa y valores por encima de 0.90 en el resto de distribuciones, alcanzando 0.96. Con $n = 50$ y $n = 100$ (ver el enlace sobre material complementario), R^2 toma valores comprendidos entre 0.91 y 0.98. Los valores más bajos se observan en las condiciones con asimetría más extrema.

En segundo lugar, la función cuadrática alcanza valores de R^2 similares a los de la función lineal, aunque ligeramente superiores en condiciones de asimetría negativa. Con $n = 50$ y $n = 100$ (ver el enlace con el material complementario), R^2 toma valores comprendidos entre 0.93 y 0.98. Los valores más bajos se observan en las condiciones con asimetrías más positivas.

En tercer lugar, el ajuste que se obtiene con la función cúbica es aún mejor que el que se obtiene con las funciones lineal y cuadrática. Con la función cúbica se obtienen valores R^2 comprendidos entre 0.96 y 0.98, aunque a costa de introducir un parámetro más y por tanto mayor complejidad. Con $n = 50$ y $n = 100$ (ver el enlace con el material complementario), el valor de R^2 está entre 0.97 y 0.995. Los valores más bajos de R^2 se observan siempre en las condiciones con mayor asimetría.

Por tanto, con las tres funciones elegidas se ha obtenido un grado de ajuste excelente. Sin embargo, desde el punto de vista aplicado, la función lineal es preferible a las otras dos porque, además de ofrecer un ajuste excelente, es la más parsimoniosa de las tres.

En la Tabla 4 se informa de los coeficientes de la función lineal. Estos coeficientes permiten estimar el porcentaje de cambios individuales a partir del tamaño del efecto: el coeficiente constante B_0 toma un valor próximo a cero en todos los casos (con valores absolutos comprendidos entre 0.01 y 2.45, y errores típicos que no sobrepasan 0.27; $p > 0.05$ en todos los casos), y el coeficiente para la pendiente B_1 se aproxima a 30 (con valores entre 28.82 y 30.78, y errores típicos nunca por debajo de 0.12). Estos resultados se encontraron de forma consistente en todas las condiciones del estudio (ver el enlace con el material complementario).

Tabla 4.

Coeficientes (y errores típicos) para el modelo de regresión lineal
($n = 25$, porcentaje de cambios basado en el estadístico *SID*)

	$\rho_{XY} = .5$		$\rho_{XY} = .7$		$\rho_{XY} = .9$	
	B_0	B_1	B_0	B_1	B_0	B_1
$As = -3, Cr = 18$	-.01 (.24)	3.08 (.11)	.48 (.25)	29.89 (.12)	1.30 (.27)	29.69 (.12)
$As = -2, Cr = 9$	1.11 (.20)	29.43 (.09)	1.09 (.21)	29.51 (.10)	1.57 (.22)	29.33 (.10)
$As = -1, Cr = 2$	1.65 (.18)	29.07 (.08)	1.76 (.18)	29.05 (.08)	2.15 (.18)	28.92 (.09)
$As = 0, Cr = 0$	1.88 (.15)	28.82 (.07)	1.89 (.15)	28.83 (.07)	1.84 (.15)	28.89 (.07)
$As = 1, Cr = 2$.42 (.17)	29.43 (.08)	.44 (.16)	29.46 (.08)	.30 (.17)	29.49 (.08)
$As = 2, Cr = 9$	-.49 (.19)	29.93 (.09)	-1.02 (.19)	30.11 (.09)	-1.12 (.20)	30.12 (.09)
$As = 3, Cr = 18$	-1.69 (.22)	30.52 (.11)	-2.45 (.23)	30.78 (.11)	-2.34 (.24)	30.68 (.12)

Las tres funciones ofrecen pronósticos muy similares. Por ejemplo, con $n = 25$, $\rho_{XY} = 0.70$ y un efecto de tamaño uno, el pronóstico (el porcentaje estimado de cambios) es del 30.7% con la función lineal, del 31.7% con la función cuadrática y del 26.9% con la función cúbica.

De acuerdo con los resultados obtenidos para la función lineal en el conjunto de condiciones simuladas podemos afirmar que: (a) cuando el tamaño del efecto es nulo ($d = 0$), el porcentaje de cambios (estimados éstos con el estadístico *SID*) se encontrará entre el 0% y el 3%; y (b) por cada punto adicional que aumente el tamaño del efecto, el porcentaje estimado de cambios se incrementará en 30 puntos. Obviamente, dado que se están pronosticando porcentajes, los pronósticos por debajo de cero y por encima de 100 deben ser reemplazados por sus respectivos límites.

Discusión

Nuestro primer objetivo era estudiar la relación entre el cambio en el centro de la distribución (tamaño del efecto) y el porcentaje de cambios individuales en un diseño pre-post de grupo único. Los diagramas de dispersión de la Figura 2 muestran la pauta de relación encontrada. Independientemente de la forma de la distribución, los resultados muestran que a medida que aumenta el tamaño del efecto, también lo hace el porcentaje de cambios.

Nuestro segundo objetivo era encontrar una función matemática capaz de modelar la relación entre el tamaño del efecto y el porcentaje de cambios. Y creemos que la hemos encontrado. Aunque los puntos de los diagramas de dispersión parecen ajustarse a una función logística o probit, el mejor ajuste se consigue con la función cúbica. Sin embargo, entre las funciones con mejor ajuste, la lineal es la más parsimoniosa (y en consecuencia la más fácil de interpretar para el profesional e investigador aplicado) y, por lo tanto, es preferible a la función cuadrática o a la cúbica. Su ajuste es excelente en todas las condiciones, con valores de R^2 que van de 0.90 a 0.96 en los escenarios menos favorables de todos los simulados ($n = 25$).

Junto con lo anterior, el hallazgo más interesante de este primer estudio es que *el valor de la pendiente de la recta de regresión es aproximadamente la misma en todas las condiciones simuladas*: en torno a 30 puntos (sus valores oscilan entre 29 y 31 puntos). Esto implica que, por cada punto que aumenta el tamaño del efecto (es decir, por cada punto que aumenta el cambio en el centro de la distribución de las diferencias pre-post en la métrica de las diferencias pre-post), la función lineal estima un incremento de 30 puntos en el porcentaje de cambios. Dicho de otra forma, a un incremento de 0.10 puntos en la diferencia estandarizada

(en la métrica de las diferencias pre-post) le corresponde un aumento de 3 puntos en el porcentaje de cambios.

Es importante señalar que estos resultados podrían haber sido parcialmente anticipados simulando la distribución de *dif* (la distribución de las diferencias pre-post) en lugar de simular las distribuciones de las puntuaciones pre y post (las distribuciones de X e Y). Si se desplaza a la derecha el centro de la distribución de *dif* y se modifica el grado de asimetría, es posible obtener el porcentaje de área bajo la curva por debajo de un determinado punto de corte. Este procedimiento permitiría simplificar sensiblemente el proceso de simulación, pero tiene algunos inconvenientes. En primer lugar, no sería posible controlar la relación entre X e Y . En consecuencia, no podríamos establecer una distribución realista para *dif*, dado que la asimetría de *dif* depende de la asimetría de X , la asimetría de Y y también de la relación XY . En segundo lugar, no podríamos determinar si la relación XY afecta o no a la relación entre el tamaño del efecto y el porcentaje de cambios (nuestros resultados sí permiten hacerlo; ver Tabla 1). En tercer lugar, no podríamos transformar d_1 (métrica de las diferencias pre-post) en d_2 (métrica de las puntuaciones pre) y, en consecuencia, no sería imposible estimar el porcentaje de cambios en aquellos contextos en los que tiende a informarse de d_2 en lugar de d_1 (contextos clínicos, meta-análisis, etc.).

Dado que el grado de relación existente entre las puntuaciones pre y post no parece alterar la relación entre la diferencia estandarizada y el porcentaje de cambios individuales, todos los resultados obtenidos para d en la métrica de las diferencias pre-post (d_1) también son válidos para d en la métrica de las puntuaciones pre (d_2). En los contextos en los que es común trabajar con d_2 , se puede transformar fácilmente d_2 en d_1 mediante $d_1 = d_2 [2(1-R_{XY})]^{1/2}$ para estimar el porcentaje de cambios asociado con cada valor de d_2 .

Para una correcta interpretación de estos resultados, debe tenerse en cuenta que si una variable Y va de cero a 100 (como lo hace el porcentaje de cambios) y la relación entre dos variables X - Y es perfecta, el valor para la pendiente de Y equivale a 100 dividido entre el rango de X . En nuestro caso, si la relación entre X (tamaño del efecto) e Y (porcentaje de cambios) fuera lineal y perfecta, la pendiente de la regresión lineal valdría $100/3.6=27.8$. Las pendientes encontradas en este estudio van de 29 a 31 porque la relación estudiada no es perfecta. Esto simplemente significa que, para hallar la pendiente correcta, es importante tener en consideración el rango de X que permita trabajar con todos los valores posibles de Y . Nuestros resultados muestran que el rango de tamaños de efecto seleccionado nos han permitido estudiar el rango completo de porcentajes de cambio individual.

La relevancia de estos hallazgos se valora mejor si se tiene en cuenta que cada vez más profesionales aplicados, y en particular psicólogos clínicos, se inclinan por evaluar la efectividad de sus tratamientos mediante el porcentaje de cambios (Ogles et al., 2001). Nuestros resultados nos llevan a la conclusión de que, cuando dicho porcentaje no es conocido, puede ser estimado a partir del tamaño del efecto grupal, que habitualmente sí lo es. En otras palabras, aunque el tamaño del efecto no permite especificar qué individuos cambian, sí informa con bastante precisión acerca del porcentaje de cambios individuales. Esta información tiene gran relevancia para evaluar la efectividad del tratamiento y es, generalmente, mucho más fácil de entender para el profesional aplicado que el tamaño del efecto.

Nuestro estudio también muestra que, contrariamente a lo que algunos autores sostienen, el porcentaje de cambios individuales no parece ser mejor indicador de la eficacia de un tratamiento que el cambio en el centro de la

distribución: en todos los escenarios simulados, ambos estadísticos son diferentes expresiones de aproximadamente la misma información.

Nuestros resultados permiten reflexionar sobre qué valores concretos hay que asociar a efectos de tamaño pequeño, medio o grande. Si consideramos los puntos de corte propuestos por Cohen (1969, pp. 24-27, 1992) para identificar efectos pequeños, medios y grandes, encontramos que, a un efecto de tamaño pequeño ($d = 0.2$) la función lineal le asigna un 8% de cambios; a un efecto de tamaño medio ($d = 0.5$) le asigna un 17% de cambios; y a un efecto de tamaño grande ($d = 0.8$) le asigna un 26% de cambios. Aunque es sabido que los puntos de corte propuestos por Cohen se deben usar con cautela, la idea de que un efecto de tamaño grande se corresponda con un 26% de cambios parece una idea poco aceptable (por decirlo de una forma suave). Desde la perspectiva del porcentaje de cambios, los puntos de corte propuestos, por ejemplo, por Ferguson (2009) parecen más razonables: $d = 0.41$ (que se corresponde con una estimación del 14.3% de cambios) para un efecto intensidad mínima; $d = 1.15$ (36.5%) para un efecto de intensidad moderada; y $d = 2.70$ (83%) para un efecto de intensidad fuerte.

Nuestro tercer objetivo era averiguar bajo qué condiciones es posible afirmar que la relación entre el tamaño del efecto y el porcentaje de cambios existe. Y los resultados obtenidos indican que esta relación se da en todas las condiciones simuladas, independientemente de la asimetría de las distribuciones pre y post, y de la correlación pre-post. Sin embargo, un examen cuidadoso de la Figura 2 revela que los cambios en la asimetría afectan visiblemente a los diagramas. Con asimetría negativa, es necesario un efecto de mayor tamaño para que el porcentaje de cambios empiece a alejarse de cero; y una vez que empieza a alejarse, lo hace muy rápido. Además, la intensidad de la relación lineal varía ligeramente entre

condiciones: en los escenarios más favorables (distribuciones normales en el pre y el post), R^2 alcanza un valor de 0.96; en los escenarios más desfavorables (asimetría negativa extrema), R^2 baja hasta 0.90. Pero esto ocurre con muestras de tamaño 25. A medida que el tamaño muestral aumenta, también lo hace el ajuste: con $n = 100$, por ejemplo, R^2 alcanza 0.98 en las condiciones más favorables y nunca queda por debajo de 0.92 en las más desfavorables.

A modo de conclusión general de este primer estudio, es posible afirmar que el cambio grupal (la estimación del cambio basada en el centro de la distribución de las diferencias pre-post) está linealmente relacionado con el porcentaje de cambios individuales (la estimación del cambio basada en la acumulación de cambios individuales). En consecuencia, es posible estimar el porcentaje aproximado de individuos que cambian después de una intervención en un contexto aplicado incluso cuando solo se conocen los indicadores grupales basados en los centros. En los escenarios que hemos simulado, la relación encontrada se puede resumir en la idea de que, para cada punto adicional que aumente el tamaño del efecto (medido como la diferencia de medias pre-post estandarizada en la métrica de las diferencias pre-post), es esperable un incremento de 30 puntos en el porcentaje de cambios individuales. Con pequeñas variaciones, este patrón es constante en todas las condiciones estudiadas.

Como ya hemos señalado, simular un diseño pre-post sin grupo control nos ha permitido llegar a conclusiones sólidas pero limitadas: no es posible, por ejemplo, evaluar la eficacia del tratamiento de forma fiable. No obstante, nos ha permitido conocer qué ocurre en un escenario simple y muy común en la investigación aplicada y en la práctica clínica. Ahora que sabemos que los indicadores basados en los centros y los basados en los individuos no son

independientes en un diseño pre-post de grupo único, pasaremos a estudiar qué pasa en otros escenarios.

Estudio 2

Relación entre el cambio basado en los centros y el cambio basado en los individuos.

II. El caso de los diseños de grupos aleatorios

En el primer estudio hemos relacionado el cambio basado en los centros (tamaño del efecto) con el cambio basado en los individuos (porcentaje de cambios) en un escenario con un solo grupo al que se le aplica una intervención y se comparan los momentos pre y post. En este segundo estudio vamos a centrarnos en la misma relación (cambio basado en los centros, cambio basado en los individuos), pero tomando como referencia un escenario completamente distinto: el diseño de dos grupos aleatorios.

Este diseño es de uso muy frecuente en estudios aplicados. En una situación bastante típica, se utiliza para comparar un grupo experimental (el grupo tratado) con un grupo control (el grupo no tratado o control). Y la valoración del efecto del tratamiento se realiza comparando ambos grupos una vez finalizado el tratamiento. Por lo tanto, estamos hablando de un escenario de dos grupos *distintos* (generalmente equivalentes o pretendidamente equivalentes sobre la base de la selección aleatoria) en los que no existe la posibilidad de valorar el

cambio individual caso a caso (tal como hemos hecho en el primer estudio) porque cada individuo únicamente aporta una puntuación.

La forma habitual (quizá exclusiva) de valorar el efecto de un tratamiento en un diseño de grupos aleatorios consiste en comparar los centros de las distribuciones de los grupos mediante estadísticos como la t de Student o la U de Mann-Whitney. Y la estrategia utilizada de forma habitual (quizá exclusiva) para cuantificar el efecto de un tratamiento en este tipo de diseños consiste en aplicar la *diferencia estandarizada d* de Cohen⁴.

No obstante, tal como ya hemos avanzado en la introducción general, existen algunos estadísticos que intentan cuantificar el tamaño del efecto comparando, no los centros de las distribuciones, sino las puntuaciones individuales. En este estudio utilizaremos el estadístico *probabilidad de superioridad (PS)*, adoptando el término propuesto por Grissom (1994; Grissom & Kim, 2012, pp. 149-175). Este estadístico informa sobre la probabilidad de que un caso seleccionado al azar del grupo tratado tenga una puntuación mayor (o menor, dependiendo de la dirección en la que apunte el efecto del tratamiento) que la de un caso seleccionado al azar del grupo no tratado o control.

Existen distintas formas de estimar la *probabilidad de superioridad* de un grupo respecto de otro. La elección de una u otra depende fundamentalmente de los supuestos que puedan hacerse sobre las varianzas de los grupos y sobre la distribución de la variable en la que se están comparando los grupos. Nosotros hemos elegido utilizar el método no paramétrico propuesto por Delaney y Vargha (2002; 2000), el cual es muy versátil porque no requiere asumir normalidad ni tampoco que los grupos tengan la misma varianza.

⁴ En un diseño de grupos aleatorios también se utiliza el coeficiente de correlación biserial-puntual para cuantificar el tamaño del efecto, pero este estadístico es equivalente a la diferencia estandarizada de Cohen. Uno puede transformarse en otro y, consecuentemente, aportan la misma información.

Estimar la *PS* mediante el método propuesto por Delaney y Vargha requiere tener información sobre las puntuaciones individuales, pues su cálculo se basa en los rangos correspondientes a esas puntuaciones. Por tanto, se trata de un estadístico que permite obtener información sobre el efecto de un tratamiento a partir de la información individual.

Objetivo

El **objetivo** general de este segundo estudio es explorar la relación existente entre la información que ofrece el cambio en los centros de las distribuciones y la información que ofrece la comparación de las puntuaciones individuales. Este objetivo general de exploración de la relación entre ambos enfoques podemos concretarlo en tres objetivos más específicos: (a) averiguar si existe relación entre el tamaño del efecto (obtenido a partir del cambio en los centros de las distribuciones) y la probabilidad de superioridad (obtenida a partir de la comparación entre las puntuaciones individuales de ambos grupos); (b) encontrar la forma de esa relación, la función matemática que la representa y su grado de ajuste; y (c) determinar bajo qué condiciones (distribución de las puntuaciones, correlación pre-post y tamaño muestral) es posible afirmar que tal relación existe.

Método

Hemos simulado datos correspondientes a una variable cuantitativa *Y* en un diseño de dos grupos aleatorios, asumiendo equivalencia entre ambos grupos en el momento pre. La variable *Y* representa la cuantificación post intervención del constructo que se desea cambiar. Por lo tanto, se espera que las medias de los grupos difieran tanto más cuanto mayor sea el efecto de la intervención.

Para definir las diferentes condiciones de la simulación, hemos manipulado cuatro criterios:

a. Tamaño muestral (n). Hemos elegido cuatro tamaños muestrales distintos (10, 25, 50 y 100 casos por grupo) para simular lo que habitualmente se consideran tamaños muestrales muy pequeños, pequeños, medianos y grandes (Crawford & Howell, 1998).

b. Forma de la distribución de Y . Hemos mantenido las siete distribuciones del primer estudio:

1. Asimetría negativa extrema: $\gamma_1 = -3$, $\gamma_2 = 18$.
2. Asimetría negativa moderada: $\gamma_1 = -2$, $\gamma_2 = 9$.
3. Asimetría negativa leve: $\gamma_1 = -1$, $\gamma_2 = 2$.
4. Normalidad: $\gamma_1 = 0$, $\gamma_2 = 0$.
5. Asimetría positiva leve: $\gamma_1 = 1$, $\gamma_2 = 2$.
6. Asimetría positiva moderada: $\gamma_1 = 2$, $\gamma_2 = 9$.
7. Asimetría positiva extrema: $\gamma_1 = 3$, $\gamma_2 = 18$.

c. Tamaño del efecto (δ). Hemos simulado el efecto de la intervención modificando la media del grupo tratado. Para esta modificación hemos utilizado efectos de trece tamaños distintos: desde 0 hasta 3.6 puntos a incrementos de 0.3 puntos. La media del grupo no tratado se ha mantenido en cero, y las desviaciones típicas de ambos grupos en 1.

La combinación de estos tres criterios ha dado lugar a $4 \times 7 \times 13 = 364$ condiciones de simulación. En cada una de estas condiciones se han simulado 500 pares de muestras (182000 pares de muestras en total). En cada par de muestras, el tamaño muestral y la forma de la distribución se han mantenido constantes; únicamente se

ha cambiado la media del grupo tratado para generar el efecto deseado. La simulación se ha llevado a cabo con el programa MatLab 2011a. El código utilizado para generar los datos puede consultarse en el Apéndice 2.

En este caso el proceso de simulación es más sencillo que en el estudio anterior. Hemos utilizado el sistema de distribuciones de Pearson (función *pearsrnd* en MatLab) para generar una matriz $\mathbf{Y} = (Y_{\text{ctrl}}, Y_{\text{tto}})$ con n parejas de puntuaciones independientes entre sí. Nótese que en este caso las dos puntuaciones de cada “pareja” no guardan relación, puesto que pertenecen a individuos distintos. En ambas variables se han impuesto los mismos valores de desviación típica ($\sigma_Y = 1$). La asimetría y curtosis de ambas se ha ajustado a los valores de γ_1 y γ_2 ya mencionados. La media de Y_{ctrl} se ha fijado a cero, mientras que la media de Y_{tto} se ha modificado sistemáticamente aplicando los valores de δ de cada condición.

Análisis de datos

En primer lugar, hemos comprobado si las propiedades de las distribuciones simuladas se correspondían con los valores especificados para generarlas. Para ello, hemos calculado, en cada condición simulada, los correspondientes estadísticos descriptivos de tendencia central, dispersión, asimetría y curtosis.

En segundo lugar, hemos calculado dos medidas del tamaño del efecto basadas en los centros. La primera de ellas es la *diferencia tipificada* d de (Cohen, 1969):

$$d = \frac{M_{\text{tto}} - M_{\text{ctrl}}}{S_{\text{combinada}}}$$

Donde

$$S_{\text{combinada}} = \sqrt{\frac{(n_{\text{tto}} - 1) S_{\text{tto}}^2 + (n_{\text{ctrl}} - 1) S_{\text{ctrl}}^2}{n_{\text{tto}} + n_{\text{ctrl}} - 2}}$$

La segunda medida de tamaño del efecto basada en los centros es el coeficiente de correlación biserial-puntual (R_{bp}):

$$R_{\text{bp}} = T / \sqrt{T^2 + n_{\text{tto}} + n_{\text{ctrl}} - 2}$$

Este coeficiente expresa la relación existente entre la variable dicotómica que define los dos grupos del diseño y la variable cuantitativa en la que se están comparando los grupos.

En tercer lugar, hemos calculado el estadístico PS para cuantificar el tamaño del efecto basándonos en la comparación por pares de cada puntuación del grupo experimental con cada puntuación del grupo control. Como ya hemos señalado, la versión de PS que hemos elegido para este trabajo (versión no paramétrica que no asume distribuciones normales ni varianzas poblacionales iguales) se puede calcular de forma bastante simple mediante las sumas de rangos utilizadas, por ejemplo, para calcular el estadístico U de Mann-Whitney, y se corresponde al área bajo la curva COR calculada con el método de Hanley y McNeil (1982). Dado que se espera que el grupo tratado tenga una media mayor que el grupo control, el valor de PS se ha estimado siempre utilizando la suma de rangos del grupo control (SUM_{ctrl}):

$$PS = [n_{\text{tto}} n_{\text{ctrl}} + n_j (n_j + 1) / 2 - SUM_{\text{ctrl}}] / (n_{\text{tto}} n_{\text{ctrl}})$$

Este estadístico toma valores comprendidos entre 0 y 1. Los valores mayores de 0.5 indican que las puntuaciones más altas pertenecen al grupo tratado. Un valor PS de, por ejemplo, 0.80 indica que, al seleccionar aleatoriamente dos

puntuaciones, una del grupo tratado y otra del grupo control, vale 0.80 la probabilidad de que la puntuación del grupo tratado sea mayor que la del grupo control.

El estadístico *PS* no es muy distinto de R_{bp} . Ambos expresan la probabilidad de que un caso de un grupo tenga mayor puntuación que un caso del otro grupo, solo que R_{bp} se basa en la métrica original de la variable y *PS* se basa en los correspondientes rangos. Esto hace que, aunque ambos estadísticos sean conceptualmente muy parecidos, no tomen el mismo valor.

Por último, hemos obtenido diagramas de dispersión para visualizar la relación subyacente entre *PS* y los estadísticos basados en los centros, y hemos ajustado varios modelos de regresión para estudiar la forma de esa relación y para cuantificar su intensidad.

Resultados

En esta sección no se incluyen los resultados de las 364 condiciones simuladas. Únicamente se incluye una parte de los resultados. No obstante, los resultados elegidos sirven como resumen y representación del conjunto completo. El resto de resultados pueden consultarse en este enlace de material complementario:

<https://copy.com/mPkovNoNQyHAXwCG>

En la Tabla 1 presentamos los estadísticos descriptivos relativos a todas las muestras simuladas con $n = 25$ y $\delta = 1.8$. Las tablas con los descriptivos sobre el resto de condiciones no aportan información adicional relevante. No obstante, pueden ser consultadas en el enlace de material complementario. A medida que el tamaño muestral aumenta, los valores obtenidos se aproximan más a los parámetros de la simulación. En las condiciones con $n = 10$ se observan resultados similares, aunque más inestables debido al reducido tamaño muestral.

En la Tabla 1 se observa que los valores de M_{tto} se aproximan al valor impuesto para δ . Las desviaciones típicas valen aproximadamente uno en ambos grupos y en todas las condiciones. Los valores de los índices de asimetría y curtosis muestran las direcciones y magnitudes esperables dado el reducido tamaño muestral. Se observa que los valores de d son ligeramente superiores a lo esperado en las condiciones con asimetrías marcadas. Este fenómeno desaparece a medida que aumenta el tamaño muestral. Los valores de R_{bp} se corresponden con los que hubiéramos obtenido a partir de δ . A pesar de que se están simulando distribuciones muy distintas, los valores del estadístico PS se mantienen entre 0.90 y 0.93; los correspondientes errores típicos están indicando que los valores de PS son más estables (varían menos) que los de los estadísticos d y R_{bp} .

Tabla 1.

Medias (desviaciones típicas) de algunos estadísticos descriptivos para los 500 pares de muestras de cada condición (con $n = 25$ y $\delta = 1.8$)

	<i>Forma de la distribución</i>						
	<i>As = -3 Cr = 18</i>	<i>As = -2 Cr = 9</i>	<i>As = -1 Cr = 2</i>	<i>As = 0 Cr = 0</i>	<i>As = 1 Cr = 2</i>	<i>As = 2 Cr = 9</i>	<i>As = 3 Cr = 18</i>
M_{ctrl}	.00 (.20)	.00 (.21)	.00 (.19)	.00 (.20)	.00 (.20)	.00 (.19)	.00 (.21)
M_{tto}	1.80 (.20)	1.80 (.20)	1.79 (.21)	1.80 (.20)	1.81 (.19)	1.81 (.20)	1.78 (.19)
S_{ctrl}	.93 (.34)	.98 (.27)	.97 (.18)	.99 (.14)	.98 (.19)	.96 (.25)	.96 (.35)
S_{tto}	.95 (.35)	.96 (.28)	1.00 (.19)	.98 (.16)	.98 (.18)	.97 (.27)	.89 (.31)
AS_{ctrl}	-1.62 (.74)	-1.15 (.69)	-.70 (.56)	.01 (.46)	.72 (.57)	1.15 (.68)	1.70 (.79)
AS_{tto}	-1.69 (.77)	-1.15 (.73)	-.73 (.59)	-.03 (.42)	.67 (.54)	1.14 (.75)	1.57 (.70)
Cr_{ctrl}	2.73 (3.40)	1.48 (2.64)	.50 (1.71)	-.17 (.77)	.55 (1.76)	1.48 (2.62)	3.08 (3.67)
Cr_{tto}	3.02 (3.61)	1.57 (2.80)	.55 (1.85)	-.28 (.69)	.36 (1.48)	1.56 (2.82)	2.52 (3.18)
R_{bp}	.69 (.11)	.68 (.09)	.67 (.08)	.67 (.07)	.68 (.07)	.68 (.09)	.69 (.11)
d	1.99 (.58)	1.91 (.47)	1.84 (.37)	1.84 (.36)	1.87 (.38)	1.91 (.44)	1.99 (.57)
PS	.93 (.04)	.92 (.04)	.91 (.04)	.90 (.04)	.91 (.04)	.92 (.04)	.93 (.04)

Las Figuras 1 y 2 contienen los diagramas de dispersión que describen la relación entre el estadístico PS (eje vertical) y los estadísticos d y R_{bp} , respectivamente (eje horizontal). Dado que la pauta de asociación subyacente se observa con más claridad cuando el tamaño muestral es mayor, las figuras se han construido con las muestras correspondientes a las condiciones con $n = 100$. Cada punto en el diagrama representa una pareja de muestras (tratamiento, control). Las pautas de asociación que se observan para $n = 100$ se observan igualmente para los demás tamaños muestrales, aunque con una mayor dispersión de los valores de PS para cada valor de d o R_{bp} .

Estos diagramas permiten apreciar con bastante claridad que el estadístico PS (basado en la comparación individual de cada caso del grupo experimental con cada caso del grupo control) no es independiente de los estadísticos d o R_{bp} (basados en la diferencia entre los centros de las distribuciones de ambos grupos). En el caso del estadístico d parece que la relación es más bien cuadrática; en el caso del estadístico R_{bp} la relación parece más bien lineal.

Figura 1.

Diagrama de dispersión con d en el eje horizontal y PS en el vertical ($n = 100$).

Cada punto del diagrama representa una estimación obtenida a partir de dos muestras (tratamiento y control)

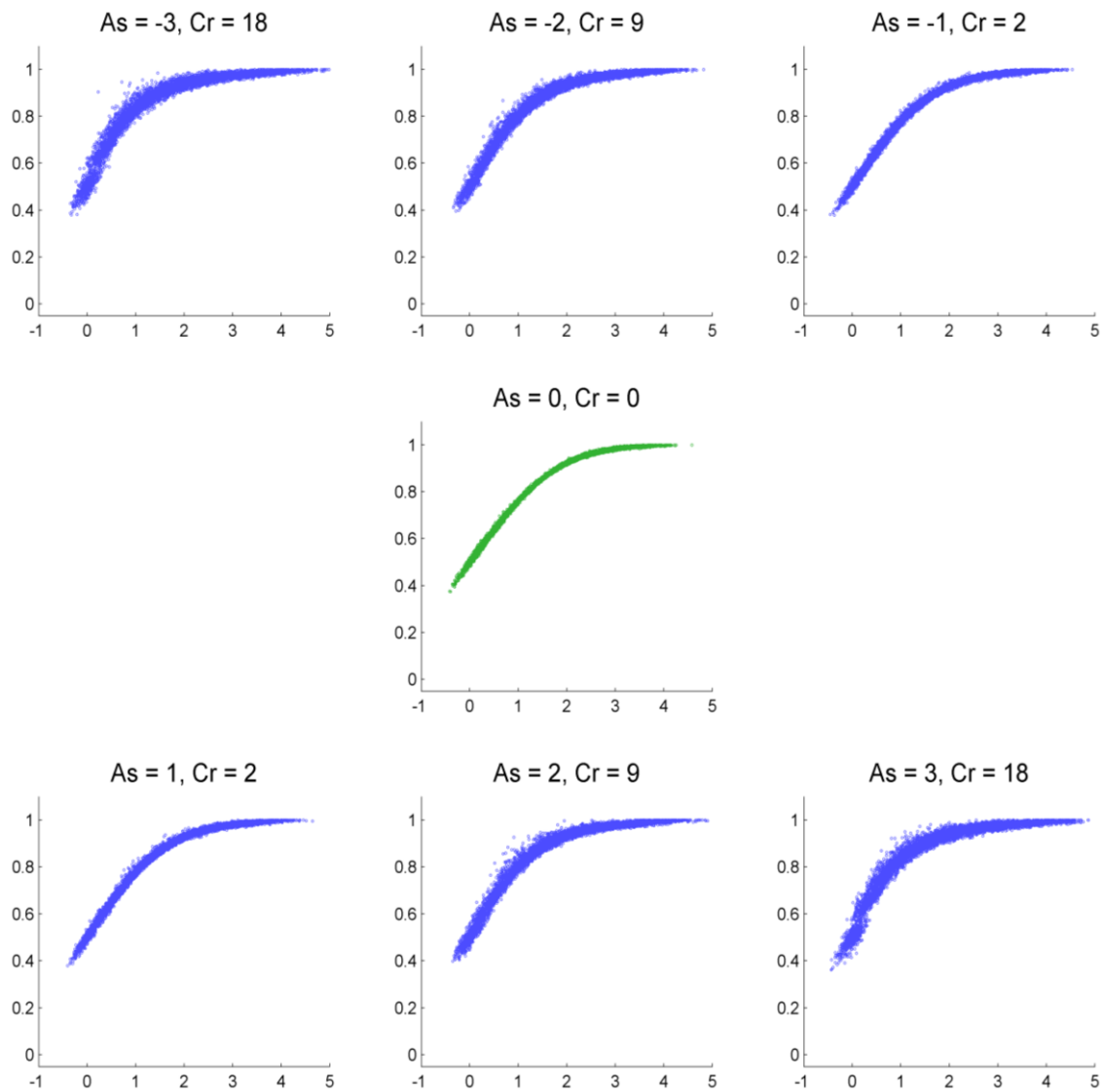
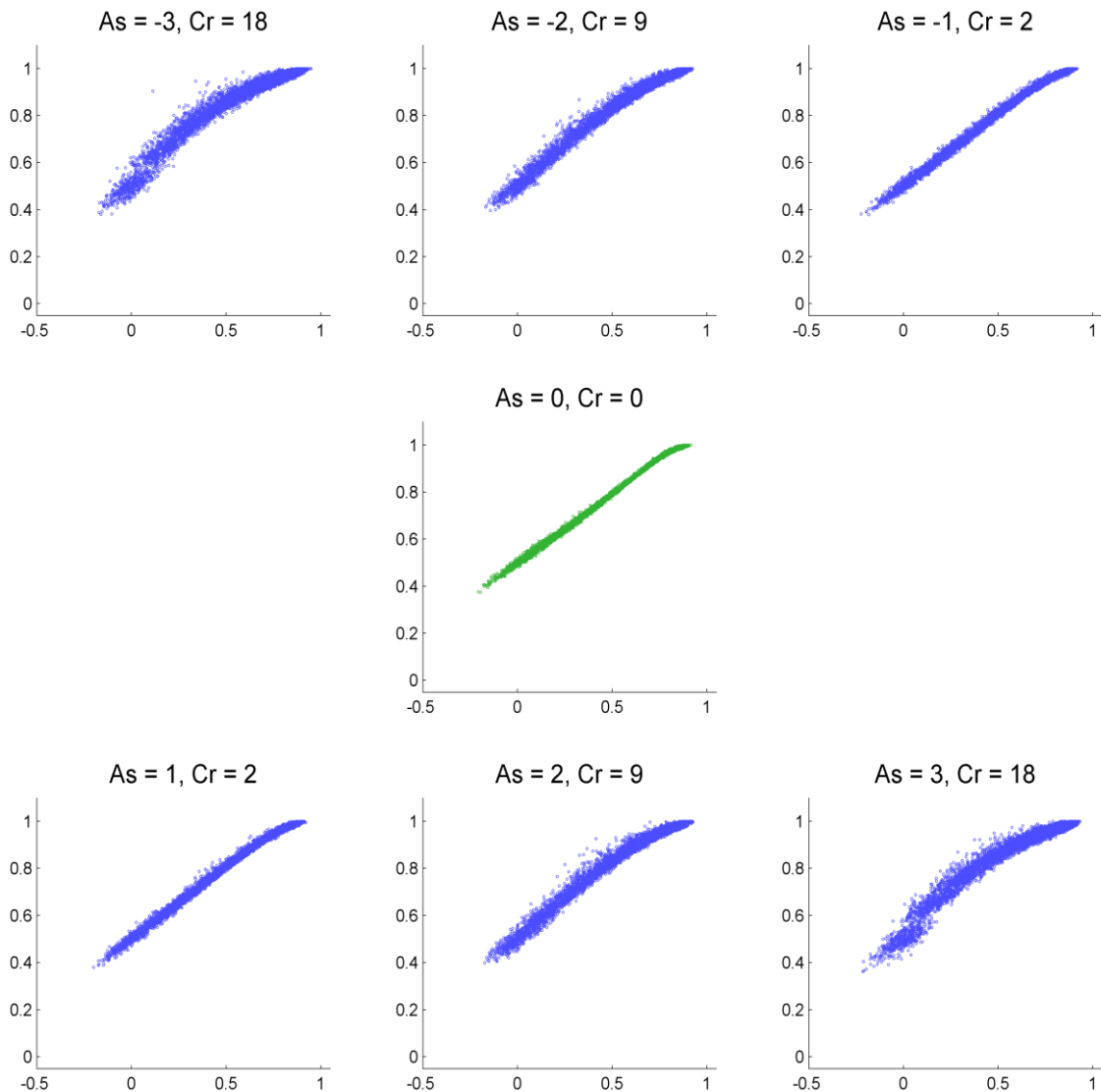


Figura 2.

Diagrama de dispersión con R_{bp} en el eje horizontal y PS en el vertical ($n = 100$).

Cada punto del diagrama representa una estimación obtenida a partir de dos muestras (tratamiento y control)



Para precisar la forma concreta de la relación hemos ajustado diferentes modelos de regresión: lineal, cuadrático, cúbico y logístico. En todos ellos hemos utilizado PS como variable dependiente y d o R_{bp} como independientes. En la Tabla 2 se ofrecen los coeficientes de regresión obtenidos con cada modelo ($n = 25$). La tabla también incluye grado de ajuste (R^2) de cada modelo.

Tabla 2a.

Resumen de los modelos de regresión con *PS* como variable dependiente y *d* como variable independiente ($n = 25$)

	<i>Modelo</i>	R^2	B_0	B_1	B_2	B_3
$As = -3$ $Cr = 18$	Lineal	.669	.689	.089		
	Cuad.	.891	.592	.225	-.029	
	Cúbico	.936	.560	.324	-.079	.006
	Logíst.	.740	1.187	.179		
$As = -2$ $Cr = 9$	Lineal	.751	.648	.107		
	Cuad.	.947	.548	.264	-.037	
	Cúbico	.966	.526	.348	-.087	.007
	Logíst.	.794	1.423	.178		
$As = -1$ $Cr = 2$	Lineal	.820	.615	.122		
	Cuad.	.979	.522	.282	-.041	
	Cúbico	.985	.510	.334	-.075	.005
	Logíst.	.828	1.610	.176		
$As = 0$ $Cr = 0$	Lineal	.852	.601	.128		
	Cuad.	.990	.510	.287	-.042	
	Cúbico	.993	.502	.322	-.064	.004
	Logíst.	.862	1.685	.176		
$As = 1$ $Cr = 2$	Lineal	.820	.616	.122		
	Cuad.	.979	.522	.283	-.041	
	Cúbico	.985	.510	.336	-.075	.005
	Logíst.	.829	1.599	.175		
$As = 2$ $Cr = 9$	Lineal	.752	.648	.107		
	Cuad.	.943	.550	.261	-.037	
	Cúbico	.963	.528	.341	-.083	.007
	Logíst.	.791	1.446	.175		
$As = 3$ $Cr = 18$	Lineal	.676	.685	.091		
	Cuad.	.894	.588	.229	-.030	
	Cúbico	.937	.557	.333	-.084	.007
	Logíst.	.734	1.188	.180		

Tabla 2b.

Resumen de los modelos de regresión con *PS* como variable dependiente
y *R_{bp}* como variable independiente ($n = 25$)

	<i>Modelo</i>	R^2	B_0	B_1	B_2	B_3
$As = -3$ $Cr = 18$	Lineal	.922	.574	.493		
	Cuad.	.943	.549	.719	-.249	
	Cúbico	.944	.546	.682	-.095	-.125
	Logíst.	.548	2.444	.001		
$As = -2$ $Cr = 9$	Lineal	.958	.538	.535		
	Cuad.	.967	.520	.696	-.177	
	Cúbico	.969	.516	.622	.127	-.251
	Logíst.	.593	2.540	.001		
$As = -1$ $Cr = 2$	Lineal	.983	.514	.563		
	Cuad.	.985	.506	.642	-.089	
	Cúbico	.987	.501	.577	.189	-.233
	Logíst.	.622	2.561	.002		
$As = 0$ $Cr = 0$	Lineal	.992	.501	.578		
	Cuad.	.992	.498	.605	-.030	
	Cúbico	.993	.495	.541	.230	-.217
	Logíst.	.657	2.741	.002		
$As = 1$ $Cr = 2$	Lineal	.983	.515	.563		
	Cuad.	.985	.505	.645	-.091	
	Cúbico	.986	.501	.575	.199	-.242
	Logíst.	.628	2.622	.002		
$As = 2$ $Cr = 9$	Lineal	.956	.539	.534		
	Cuad.	.964	.521	.693	-.175	
	Cúbico	.966	.517	.624	.110	-.235
	Logíst.	.589	2.559	.001		
$As = 3$ $Cr = 18$	Lineal	.920	.572	.495		
	Cuad.	.942	.547	.723	-.251	
	Cúbico	.942	.544	.691	-.110	-.117
	Logíst.	.549	2.394	.001		

Como se observa en la Tabla 2, los modelos de regresión obtienen, por lo general, índices de ajuste aceptables incluso con un tamaño muestral de solo 25 casos. A medida que el tamaño muestral aumenta, el ajuste de los modelos va mejorando. Los resultados obtenidos al ajustar estos modelos de regresión con el resto de tamaños muestrales pueden consultarse en el enlace de material complementario.

El modelo que mejor ajusta en todas las condiciones es el cúbico. No obstante, cuando se utiliza d como variable independiente, el modelo de regresión cuadrática permite obtener un grado de ajuste muy parecidos en todas las condiciones ($R^2 = 0.891$ en el peor caso, $R^2 = 0.990$ con distribuciones normales). Y cuando se utiliza R_{bp} como variable independiente, el modelo lineal consigue un ajuste excelente ($R^2 = 0.922$ en el peor caso, $R^2 = 0.992$ con distribuciones normales).

Dado que, en ausencia de diferencias en el ajuste, los modelos parsimoniosos son preferibles a los más complejos, consideramos que: (1) para modelar la relación entre d y PS la función idónea es la cuadrática; y (2) para modelar la relación entre R_{bp} y PS la función idónea es la línea.

Discusión

Los resultados de este segundo estudio aportan información precisa sobre los tres objetivos planteados al principio. En primer lugar, puede afirmarse que, en los diseños en los que dos grupos aleatorios se comparan en una variable cuantitativa, ciertas estimaciones del tamaño del efecto basadas en información individual, más concretamente, las estimaciones basadas en el estadístico PS propuesto por Delaney y Vargha (Delaney & Vargha, 2002; Grissom & Kim, 2012, p. 159; Vargha & Delaney, 2000), están estrechamente relacionadas con ciertas estimaciones del tamaño del efecto basadas en información grupal, más concretamente, con las estimaciones obtenidas aplicando dos de los métodos más utilizados para cuantificar el tamaño del efecto en este tipo de diseños: la diferencia estandarizada d de Cohen y el coeficiente de correlación biserial-puntual R_{bp} . Estos resultados son congruentes con los de nuestro primer estudio, que revelaron un patrón similar en un diseño pre-post de grupo único.

En segundo lugar, entre los distintos modelos de regresión que hemos ajustado, el cuadrático parece ser el modelo más adecuado para dar cuenta de la relación entre PS y d ; y el lineal parece ser el modelo más adecuado para dar cuenta de la relación entre PS y R_{bp} . El grado de ajuste que se obtiene en ambos casos es excelente. Los coeficientes de determinación R^2 alcanzan valores muy altos (incluso con 25 casos por grupo), lo que indica que las predicciones que hacen los respectivos modelos de regresión son bastante precisas.

En tercer lugar, es destacable que los coeficientes de los modelos de regresión no cambian sustancialmente cuando cambia la forma de la distribución de las puntuaciones. Por tanto, parece que la forma de las distribuciones no está afectando de forma sensible a las relaciones encontradas. En principio, parece que puede asumirse que los patrones de asociación encontrados son independientes de las distribuciones simuladas.

Los resultados de este estudio tienen algunas implicaciones de cierta relevancia de cara a la evaluación de intervenciones en contextos aplicados. Como ya hemos señalado en el primer estudio, Cohen, además de proponer el estadístico d (Cohen, 1969, 1988) propuso una especie de *guía* para poder interpretarlo (Cohen, 1988, 1992). Esta guía incluye tres puntos de corte que sirven de referencia para clasificar un efecto como de tamaño pequeño ($d = 0.2$), mediano ($d = 0.5$) o grande ($d = 0.8$). Si bien estos puntos de corte son del todo arbitrarios, la insistencia de Cohen (1992) en proponerlos como referentes para la interpretación (junto con, probablemente, la necesidad de los investigadores aplicados de tener algo a lo que agarrarse para poder valorar los efectos que estudian) ha terminado por convertirlos en una especie de “regla mágica” para la mayoría de los investigadores.

Sin embargo, los resultados de este estudio (como también los del primero) nos hacen dudar seriamente de que dichos puntos de corte estén bien elegidos. En

la Tabla 3 se muestran algunos pronósticos para *PS* calculados con nuestro modelo de regresión cuadrática.

Tabla 3.

Pronósticos del modelo de regresión cuadrática para *PS* con distribuciones normales: $PS_{\text{pronosticado}} = 0.510 + 0.287(d) - 0.042(d^2)$.

<i>d observado</i>	.0	.2	.5	.8	1.0	1.5	1.7	2.0	2.5	3.0	3.5
<i>PS pronosticado</i>	.51	.57	.64	.71	.76	.85	.88	.92	.97	.99	1.0

Siguiendo los criterios de Cohen, un efecto pequeño ($d = 0.2$) se corresponde con una probabilidad de superioridad de 0.57 ($PS = 0.57$), un efecto mediano ($d = 0.5$) se corresponde con $PS = 0.64$; y un efecto grande ($d = 0.8$) se corresponde con $PS = 0.71$. Estos valores de *PS* indican que la probabilidad de que un caso seleccionado al azar del grupo tratado tenga una puntuación mayor que la de un caso seleccionado al azar del grupo control vale 0.57, 0.64 y 0.71, respectivamente.

Para poder interpretar este dato es necesario tener en cuenta que a la ausencia de efecto (al efecto nulo) le corresponde una *PS* de 0.5. Es decir, si el grupo tratado no difiere del control la probabilidad de que un caso cualquiera del grupo tratado tenga una puntuación *mayor* que la de un caso cualquiera del grupo control es la misma que la probabilidad de que un caso cualquiera del grupo tratado tenga una puntuación *menor* que la de un caso cualquiera del grupo control. En este escenario, 0.57, 0.64 y 0.71 parecen probabilidades muy bajas para identificar efectos de tamaño pequeño, medio y grande. Los puntos de corte propuestos por Ferguson (2009) podrían constituir una alternativa aceptable a los de Cohen. Al traducirlos a *PS*, un efecto de tamaño *mínimo* ($d = 0.41$) se corresponde con $PS = 0.62$; un efecto de tamaño *moderado* ($d = 1.15$) se corresponde con $PS = 0.78$; y un efecto de tamaño *fuerte* ($d = 2.7$) se corresponde con $PS = 0.98$.

Para valorar correctamente estas correspondencias es importante tener en cuenta que $PS = 0.98$ no significa que el 98% de los casos del grupo tratado tengan puntuaciones mayores cualquier caso del grupo control. Significa que, si un efecto es de tamaño grande ($d = 2.7$ en la propuesta de Ferguson) es improbable (probabilidad 0.02) que un sujeto seleccionado al azar del grupo control tenga un puntuación mayor que un sujeto seleccionado al azar del grupo tratado.

Estudio 3

Relación entre el cambio basado en los centros y el cambio basado en los individuos.

III. El caso de los diseños pre-post con grupos aleatorios

Al igual que en los dos estudios anteriores, en éste seguiremos estudiando la relación entre la información que ofrecen los métodos basados en el cambio grupal y la que ofrecen los métodos basados en el cambio individual.

En el primer estudio hemos analizado esta relación en el contexto definido por un diseño pre-post de grupo único; en el segundo estudio hemos analizado exactamente la misma relación en un diseño con grupos aleatorios. Ambos diseños son muy utilizados en el ámbito de las ciencias de la salud, pero ambos adolecen de importantes debilidades. La principal debilidad del diseño pre-post de grupo único es que, al no existir grupo control, no es posible separar el efecto de la intervención del efecto de otros factores no controlados (como, por ejemplo, la evolución de los sujetos con el paso del tiempo). La principal debilidad del diseño de grupos aleatorios es que la falta de seguimiento impide valorar si el efecto estudiado perdura en el tiempo; además, para la correcta valoración del efecto del tratamiento, es necesario asumir que los grupos utilizados son equivalentes (pues

no hay posibilidad de comparar lo que ocurre antes del tratamiento con lo que ocurre después).

En este tercer y último estudio vamos a extender nuestro análisis a la relación entre los enfoques individual y grupal a un *diseño pre-post con grupos aleatorios*. Se trata de un diseño que, al incluir grupo control, supera la debilidad del diseño pre-post de grupo único; y, al hacer seguimiento, supera las debilidades del diseño de grupos aleatorios (Feingold, 2009; Shadish et al., 2002, pp. 108-110).

Por tanto, nuestro estudio de la relación entre diferentes métodos de evaluación del cambio va estar ahora centrado en un diseño con dos grupos aleatorios (tratamiento, control) sometidos a dos mediciones distintas (antes y después de la intervención). El hecho de contar con dos grupos y dos mediciones permite realizar una valoración más fiable y eficiente del cambio, pues a diferencia de lo que ocurría en los otros dos diseños, ahora es posible comparar lo que ocurre en el momento pre (la diferencia entre los grupos en el momento pre) con lo que ocurre en el momento post (la diferencia entre grupos los en el momento post)

Objetivo

El **objetivo general** de este estudio es explorar la relación entre diferentes métodos de evaluación del cambio en diseños pre-post con grupos aleatorios. Más concretamente, la relación entre la información que se obtiene con el enfoque basado en los centros de las distribuciones y la que se obtiene con el enfoque basado en las puntuaciones individuales.

Este propósito general de exploración de la relación entre ambos enfoques podemos concretarlo en tres objetivos más específicos: (a) averiguar si existe relación entre el tamaño del efecto (obtenido a partir del cambio en los centros de las distribuciones) y el porcentaje de cambios (obtenido a partir de la acumulación

de cambios individuales); (b) encontrar la forma de esa relación, la función matemática que la representa y su grado de ajuste; y (c) determinar bajo qué condiciones (distribución de las puntuaciones, correlación pre-post y tamaño muestral) es posible afirmar que tal relación existe.

Método

Para cubrir nuestros objetivos hemos simulado datos correspondientes a un diseño pre-post con dos grupos aleatorios (tratamiento, control). Las condiciones de la simulación se han definido manipulando cuatro criterios:

- a. *Tamaño del efecto* (δ). El tamaño del efecto impuesto a cada grupo en cada condición simulada lo hemos controlado mediante la diferencia estandarizada δ de Cohen. Para obtener este estadístico hemos tipificado el cambio entre el pre y el post en cada grupo, es decir, la media de las diferencias pre-post de cada grupo (μ_{dif}), utilizando la desviación típica de las correspondientes diferencias pre-post (σ_{dif}):

$$\delta_{\text{ctrl}} = \mu_{\text{dif}(\text{ctrl})} / \sigma_{\text{dif}(\text{ctrl})}$$

$$\delta_{\text{tto}} = \mu_{\text{dif}(\text{tto})} / \sigma_{\text{dif}(\text{tto})}$$

Las puntuaciones del grupo control se han generado asumiendo que entre los momentos pre y post no existen diferencias, es decir, imponiendo $\delta_{\text{ctrl}} = 0$. Las puntuaciones del grupo tratado se han generado imponiendo efectos de 13 tamaños distintos, es decir, imponiendo 13 valores distintos para δ_{tto} : de 0 a 3.6 puntos, a intervalos de 0.3 puntos. Así, un efecto de tamaño 1.2 indica que la media de las diferencias del grupo tratado es igual a 1.2 veces la desviación típica de las diferencias pre-post en ese grupo. Como ya hemos podido comprobar en el primer estudio, este rango de valores (de 0 a 3.6) para el estadístico δ asegura que el porcentaje de cambios individuales del grupo

tratado pueda tomar valores en todo su rango (de 0% a 100%). Estamos asumiendo que el efecto de la intervención hace aumentar las puntuaciones. Y para calcular las diferencias pre-post hemos restado la puntuación pre a la puntuación post. En consecuencia, estamos trabajando con efectos positivos.

- b. *Tamaño muestral (n)*. Hemos elegido cuatro tamaños muestrales: 10, 25, 50 y 100 *por grupo* para simular lo que habitualmente se consideran tamaños muestrales muy pequeños, pequeños, medianos y grandes (Crawford & Howell, 1998). Nótese que estos tamaños muestrales se refieren a cada grupo (no a la suma de ambos). Es decir, cuando hablamos de tamaño $n = 25$ quiere decir que hay 25 individuos en el grupo control y otros 25 en el grupo tratado.
- c. *Correlación pre-post (ρ_{XY})*: 0.5, 0.7 y 0.9. Hemos elegido estos valores para simular fiabilidades o correlaciones pre-post moderadas, medias y altas (Nunnally & Bernstein, 1994, pp. 264-265; Pedhazur & Schmelkin, 1991, pp. 109-110). Para cuantificar el grado de relación hemos utilizado el coeficiente de correlación de Pearson.
- d. *Forma de las distribuciones*. Hemos mantenido las siete distribuciones del primer estudio:
 1. Asimetría negativa extrema: $\gamma_1 = -3, \gamma_2 = 18$.
 2. Asimetría negativa moderada: $\gamma_1 = -2, \gamma_2 = 9$.
 3. Asimetría negativa leve: $\gamma_1 = -1, \gamma_2 = 2$.
 4. Normalidad: $\gamma_1 = 0, \gamma_2 = .$
 5. Asimetría positiva leve: $\gamma_1 = 1, \gamma_2 = 2$.
 6. Asimetría positiva moderada: $\gamma_1 = 2, \gamma_2 = 9$.
 7. Asimetría positiva extrema: $\gamma_1 = 3, \gamma_2 = 18$.

Según Blanca *et al.* (2013), menos del 5% de las muestras reales tienen distribuciones más extremas.

Combinando estos cuatro criterios hemos creado $13 \times 4 \times 3 \times 7 = 1092$ condiciones de simulación distintas. En cada una de estas condiciones se han generado 500 muestras (546000 muestras en total). En cada réplica, el tamaño muestral, la correlación pre-post y el grado de asimetría y curtosis fueron idénticos para cada pareja de grupos tratamiento-control. Los dos grupos de cada réplica únicamente diferían en el tamaño del efecto impuesto.

Para realizar la simulación hemos utilizado el programa MatLab 2011a. El procedimiento seguido es similar al ya explicado en el estudio 1, pero en lugar de una sola matriz \mathbf{X} , se han generado dos: $\mathbf{X}_1 = (X_1^*, Y_1^*)$ y $\mathbf{X}_2 = (X_2^*, Y_2^*)$, cada una con n parejas de puntuaciones no correlacionadas. Estas dos matrices contienen las puntuaciones del grupo control (\mathbf{X}_1) y del grupo tratado (\mathbf{X}_2).

En primer lugar se han generado las puntuaciones de las matrices \mathbf{X}_1 y \mathbf{X}_2 utilizando el sistema de distribuciones de Pearson (función *pearsrnd* en MatLab). Las cuatro variables generadas (dos por matriz) tenían la misma media, desviación típica, asimetría y curtosis. Todas las medias se fijaron en cero y todas las desviaciones típicas en uno. El grado de asimetría y de curtosis se fue modificando sistemáticamente de acuerdo a los valores γ_1 y γ_2 detallados anteriormente.

En segundo lugar hemos fijado el valor de la correlación entre las dos variables de cada matriz aplicando la descomposición de Cholesky (función *cholcov* en MatLab) de la matriz de correlaciones \mathbf{R} correspondiente al valor elegido para la correlación ρ_{XY} en cada condición simulada. Las matrices resultantes, $\mathbf{M}_1 = (X_1, Y_1)$ y $\mathbf{M}_2 = (X_2, Y_2)$, contenían cada una dos variables ($X = \text{pre}$; $Y = \text{post}$) con valores de asimetría, curtosis y correlación entre ellas similares a los especificados.

Finalmente hemos procedido a modificar Y_2 en la matriz del grupo tratado (no en la del grupo control) para hacer que su media se aproximase a la media

elegida para cada condición simulada. Para esto, a cada valor individual de Y_2 se le ha sumado, en cada muestra, el resultado de multiplicar la desviación típica de las diferencias pre-post del grupo tratado por el tamaño del efecto correspondiente a la condición simulada. El código de la simulación se puede consultar en el Apéndice 2.

Análisis de datos

En primer lugar, tras generar las 546000 muestras, hemos comprobado si las propiedades de las distribuciones simuladas eran congruentes con los valores impuestos para generarlas. Para ello, hemos calculado, en cada condición simulada, estadísticos descriptivos de tendencia central, de dispersión, de asimetría y de curtosis. También hemos calculado la correlación pre-post y hemos comprobado si el tamaño de los efectos observados se correspondía con los tamaños impuestos en cada condición simulada.

En segundo lugar, hemos calculado, en cada muestra simulada, el tamaño del efecto basado en los centros de las distribuciones. Puesto que estamos utilizando un diseño mixto con un factor inter-sujetos (*grupo* o factor *A*) y otro factor intra-sujetos (*momento pre-post* o factor *B*), hemos cuantificado el cambio basado en los centros a partir del efecto de la interacción *AB*. En concreto, hemos utilizado las siguientes dos medidas del tamaño del efecto:

- a.* El estadístico ω^2 asociado al efecto de la interacción *AB*, el cual captura el efecto de la intervención mediante la comparación del cambio pre-post del grupo tratado con el cambio pre-post del grupo control (Kirk, 2013, pp. 77-124, 134-138). En un diseño pre-post con grupos aleatorios, ω^2 puede estimarse fácilmente utilizando la información que ofrece el modelo de ANOVA de dos factores con medidas repetidas en uno:

$$\hat{\omega}^2 = \frac{gl_{\text{interacción}} (F_{\text{interacción}} - 1)}{gl_{\text{interacción}} (F_{\text{interacción}} - 1) + N}$$

Debe tenerse en cuenta que N no se refiere al número de casos sino al número de puntuaciones. Por tanto, en un diseño pre-post con dos grupos aleatorios y 25 casos por grupo, N vale $2 \times 2 \times 25 = 100$.

- b. Una versión de la diferencia estandarizada d de Cohen (1969) propuesta por Morris (2008) y Grissom y Kim (2012, pp. 90-92) para analizar el efecto de la interacción en el contexto del meta-análisis, en lo que ellos llaman *pretest-posttest-control group designs (PPC)*:

$$d_{PPC} = \frac{(M_{\text{post; tto}} - M_{\text{post; ctrl}}) - (M_{\text{pre; tto}} - M_{\text{pre; ctrl}})}{S_{\text{pre}}}$$

donde S_{pre} , la desviación típica combinada de los dos grupos en el momento pre, puede calcularse (con $n_{\text{tto}} = n_{\text{ctrl}}$) de la siguiente manera:

$$S_{\text{pre}} = \sqrt{(S_{\text{pre; tto}}^2 + S_{\text{pre; ctrl}}^2) / 2}$$

En tercer lugar, hemos utilizado índices de cambio individual para obtener una estimación del porcentaje de cambios individuales. Para ello, primero hemos aplicado los dos índices de cambio fiable detallados en el Estudio 1: la *diferencia individual tipificada* (*standardized individual difference, SID*) y el *índice de cambio fiable* (*reliable change index, RCI*). Ambos índices se basan en la tipificación de las diferencias individuales y con ambos se considera que una diferencia es fiable cuando cae por encima (o por debajo) de un determinado punto de corte. Cuando la distribución de las diferencias pre-post es normal, el 95% de los valores de *SID* y *RCI* se encuentran entre ± 1.96 desviaciones típicas. En consecuencia, hemos considerado que un caso ha cambiado (empeorado o mejorado) cuando su

correspondiente diferencia pre-post tipificada se encontraba por encima o por debajo de dichos puntos de corte. Debe tenerse en cuenta que, dado que solo hemos trabajado con efectos positivos, es esperable que el porcentaje de casos que empeoran sea pequeño (y tendente a cero cuanto mayor sea el efecto de la intervención). En este escenario podría ser adecuado el uso de contrastes unilaterales derechos. Sin embargo, dado que nuestro procedimiento está diseñado para comparar la eficacia de dos tratamientos distintos en escenarios reales, es importante tener en cuenta también el porcentaje de casos que empeoran, y no solo los que mejoran. Consecuentemente, hemos utilizado contrastes bilaterales para decidir qué valores de *SID* y *RCI* se declaran significativos.

Después de decidir qué individuos habían experimentado un cambio fiable, hemos calculado el porcentaje de casos que mejoraron y empeoraron en cada muestra. A continuación hemos restado el porcentaje de casos que ha empeorado (P^-) al porcentaje de casos que ha mejorado (P^+) en cada uno de los dos grupos, y después hemos restado el resultado de dicha operación en el grupo control (*ctrl*) al resultado de esa misma operación en el grupo tratado (*tto*) para llegar al *porcentaje neto de mejorías* atribuible al tratamiento (P_{NM}). Este porcentaje neto constituye un indicador directo de la eficacia de la intervención y cumple con el requisito de estar basado en la información individual de todos los casos de ambos grupos:

$$P_{NM} = (P_{tto}^+ - P_{tto}^-) - (P_{ctrl}^+ - P_{ctrl}^-)$$

Norman, Sridhar, Guyatt y Walter (2001), inspirados en Guyatt, Juniper, Walter, Griffith y Goldstein (1998), han propuesto una forma alternativa de calcular P_{NM} :

$$P_{NM} = (P_{tto}^+ - P_{tto}^-) - (P_{ctrl}^+ - P_{ctrl}^-) - P_{tto}^- P_{ctrl}^+ - P_{tto}^+ P_{ctrl}^-$$

Sin embargo, como los resultados obtenidos con ambas ecuaciones son casi idénticos, nos limitaremos a informar de los obtenidos con la primera.

Finalmente, hemos estudiado la relación entre las estimaciones del cambio basadas en los centros y las estimaciones del cambio basadas en los individuos. Para ello, hemos creado varios diagramas de dispersión y hemos ajustado varios modelos de regresión.

Resultados

Al igual que en los dos estudios anteriores, en esta sección solamente incluiremos los resultados de algunas de las condiciones simuladas. Los resultados de las condiciones no informadas aquí no aportan información relevante adicional para comprender la relación entre el cambio basado en los centros y el cambio basado en los individuos. No obstante, todos estos resultados se pueden consultar en los enlaces web incluidos en el siguiente enlace:

<https://copy.com/mPkovNoNQyHxwCG>

En la Tabla 1 se proporciona un resumen descriptivo de los datos simulados en cada una de las siete condiciones con $n = 25$, $\rho_{XY} = 0.7$ y $\delta_{tto} = 1.2$. La tabla incluye las medias (M) y las desviaciones típicas (S) de las puntuaciones pre, de las puntuaciones post y de las diferencias post-pre (dif). También incluye los índices de asimetría y curtosis, el coeficiente de correlación de Pearson entre las puntuaciones del pre y del post (R_{XY}) y el tamaño del efecto observado (d). Toda esta información se ofrece tanto para el grupo control (*ctrl*) como para el grupo tratado (*tto*). Adicionalmente, se incluyen dos estimaciones del tamaño del efecto de la interacción AB: d_{PPC} y $\hat{\omega}^2$. Los resultados relativos al resto de condiciones simuladas pueden consultarse en el enlace de material complementario.

Tabla 1.

Estadísticos descriptivos correspondientes a las siete condiciones simuladas con $n = 25$, $\rho_{XY} = 0.7$ y $\delta_{tto} = 1.2$. Cada casilla incluye la media (desv. típ.) de las 500 muestras simuladas en cada condición

	$As = -3$ $Cr = 18$	$As = -2$ $Cr = 9$	$As = -1$ $Cr = 2$	$As = 0$ $Cr = 0$	$As = 1$ $Cr = 2$	$As = 2$ $Cr = 9$	$As = 3$ $Cr = 18$
$M_{pre,ctrl}$.00 (.21)	.02 (.20)	.01 (.20)	.01 (.21)	.00 (.20)	.00 (.20)	.00 (.20)
$M_{pre,tto}$.00 (.19)	.01 (.20)	-.01 (.19)	.00 (.19)	.00 (.20)	.01 (.20)	-.01 (.19)
$M_{post,ctrl}$.00 (.20)	.01 (.21)	.00 (.20)	.00 (.21)	-.01 (.20)	.00 (.20)	-.01 (.20)
$M_{post,tto}$.89 (.24)	.90 (.21)	.90 (.21)	.92 (.24)	.92 (.31)	.92 (.38)	.88 (.41)
$M_{dif,ctrl}$.00 (.16)	.00 (.15)	-.01 (.16)	-.01 (.16)	-.01 (.16)	.00 (.16)	-.01 (.15)
$M_{dif,tto}$.89 (.24)	.89 (.18)	.91 (.18)	.91 (.20)	.92 (.26)	.91 (.35)	.89 (.38)
$S_{pre,ctrl}$.94 (.37)	.96 (.25)	.97 (.18)	1.00 (.14)	.99 (.19)	.97 (.27)	.94 (.32)
$S_{pre,tto}$.93 (.32)	.94 (.25)	.99 (.18)	.99 (.14)	.98 (.19)	.97 (.28)	.93 (.31)
$S_{post,ctrl}$.96 (.29)	.97 (.22)	.98 (.17)	1.00 (.14)	.99 (.17)	.97 (.23)	.95 (.27)
$S_{post,tto}$.97 (.28)	.96 (.20)	.99 (.17)	.99 (.14)	.98 (.17)	.98 (.24)	.96 (.25)
$S_{dif,ctrl}$.73 (.24)	.74 (.19)	.76 (.15)	.78 (.11)	.75 (.14)	.76 (.21)	.73 (.24)
$S_{dif,tto}$.75 (.25)	.75 (.18)	.75 (.15)	.77 (.11)	.76 (.14)	.76 (.20)	.73 (.22)
$AS_{pre,ctrl}$	-1.62 (.75)	-1.18 (.66)	-.66 (.55)	.02 (.43)	.70 (.56)	1.16 (.69)	1.65 (.76)
$AS_{pre,tto}$	-1.61 (.74)	-1.12 (.69)	-.70 (.55)	.02 (.43)	.70 (.58)	1.17 (.71)	1.62 (.74)
$AS_{post,ctrl}$	-1.29 (.70)	-.90 (.64)	-.45 (.49)	.00 (.43)	.49 (.51)	.88 (.65)	1.28 (.67)
$AS_{post,tto}$	-1.30 (.69)	-.85 (.61)	-.51 (.51)	-.01 (.43)	.53 (.50)	.93 (.65)	1.28 (.64)
$AS_{dif,ctrl}$	-1.06 (.96)	-.82 (.76)	-.51 (.53)	-.04 (.41)	.46 (.54)	.82 (.80)	1.09 (.94)
$AS_{dif,tto}$	-1.14 (.95)	-.79 (.73)	-.46 (.57)	-.01 (.41)	.52 (.54)	.77 (.79)	1.13 (.90)
$Cr_{pre,ctrl}$	2.75 (3.39)	1.59 (2.53)	.38 (1.52)	-.22 (.70)	.48 (1.59)	1.52 (2.69)	2.87 (3.52)
$Cr_{pre,tto}$	2.69 (3.41)	1.45 (2.69)	.47 (1.61)	-.24 (.71)	.49 (1.70)	1.61 (2.86)	2.71 (3.44)
$Cr_{post,ctrl}$	1.78 (2.87)	.91 (2.24)	.05 (1.16)	-.21 (.72)	.11 (1.29)	.87 (2.19)	1.71 (2.75)
$Cr_{post,tto}$	1.83 (2.89)	.80 (1.93)	.17 (1.32)	-.23 (.69)	.15 (1.20)	1.02 (2.34)	1.68 (2.56)
$Cr_{dif,ctrl}$	2.27 (2.98)	1.35 (2.40)	.26 (1.27)	-.25 (.73)	.28 (1.21)	1.32 (2.63)	2.33 (3.09)
$Cr_{dif,tto}$	2.44 (3.01)	1.13 (2.19)	.26 (1.41)	-.26 (.69)	.33 (1.31)	1.16 (2.36)	2.26 (3.07)
$R_{XY,ctrl}$.68 (.18)	.69 (.15)	.69 (.12)	.69 (.11)	.70 (.12)	.68 (.15)	.69 (.17)
$R_{XY,tto}$.67 (.19)	.68 (.15)	.70 (.12)	.69 (.11)	.69 (.12)	.68 (.16)	.68 (.17)
d_{ctrl}	.05 (.24)	.02 (.21)	.01 (.22)	-.02 (.21)	-.03 (.22)	-.02 (.22)	-.06 (.22)
d_{tto}	1.23 (.22)	1.22 (.21)	1.23 (.21)	1.19 (.20)	1.19 (.21)	1.17 (.22)	1.18 (.22)
d_{ppc}	.98 (.38)	.96 (.29)	.95 (.27)	.94 (.27)	.95 (.34)	.95 (.43)	.99 (.50)
$\hat{\omega}^2$.14 (.05)	.15 (.05)	.15 (.06)	.15 (.06)	.15 (.08)	.15 (.08)	.16 (.09)

A medida que el tamaño muestral aumenta, los datos simulados se asemejan más a los valores de referencia utilizados para generarlos. Por lo tanto, los resultados obtenidos con $n = 25$ dan una idea precisa de la calidad de la simulación. Las condiciones con mayores tamaños muestrales ofrecieron menores desviaciones típicas para cada uno de los estadísticos. Los resultados obtenidos con $n = 10$ también son similares a los valores de referencia, aunque se alejan más que los tamaños muestrales mayores. Dado que son más difíciles de interpretar por su inestabilidad, hemos elegido presentar aquí los resultados obtenidos con $n = 25$. Consideramos que ese tamaño muestral es, de todos los elegidos, el más común en diferentes contextos aplicados en el ámbito de las ciencias sociales y de la salud. Además, con ese tamaño muestral empiezan a desaparecer los problemas del estadístico F en condiciones de no-normalidad.

Los resultados de la Tabla 1 indican que las propiedades de las distribuciones generadas se corresponden con las propiedades impuestas: las desviaciones típicas de las puntuaciones pre y post toman valores próximos a 1; R_{XY} toma valores próximos a 0.7; las desviaciones típicas de las diferencias (S_{dif}) toman valores próximos a

$$\sigma_{dif} = [\sigma_{pre}^2 + \sigma_{post}^2 - 2\sigma_{pre} \sigma_{post} \rho_{XY}]^{1/2} = [1 + 1 - 2(1)(1)(0.7)]^{1/2} = 0.775$$

El tamaño del efecto en el grupo tratado (d_{tto}) toma un valor próximo a 1.2; el tamaño del efecto en el grupo control (d_{ctrl}) toma un valor próximo a cero; las medias pre y post del grupo control ($M_{pre;ctrl}$ y $M_{post;ctrl}$) y la media pre del grupo tratado ($M_{pre;tto}$) toman valores próximos a cero; y la media post del grupo tratado toma un valor próximo a 0.93 ($\mu_{post;tto} = \mu_{pre;tto} + \delta_{tto} \sigma_{dif} = 0 + 1.2(0.775) = 0.93$). Los índices de asimetría y curtosis estuvieron por debajo de los valores especificados, pero tomaron los valores esperables para este tamaño muestral. Respecto a las estimaciones del tamaño del efecto, $\hat{\omega}^2$ ha tomado un valor próximo

a 0.15 y d_{PPC} un valor próximo a 0.93; por tanto, los valores obtenidos fueron, en ambos casos, muy parecidos a los esperados.

Nótese que los valores que cabe esperar que tomen $\hat{\omega}^2$ y d_{PPC} se deducen del valor impuesto para el tamaño del efecto del grupo tratado δ_{tto} . Al dividir el valor estimado de δ_{tto} (es decir, d_{tto}) entre la raíz cuadrada de la suma de los inversos de los tamaños muestrales se obtiene el estadístico t de Student que permite comparar la media de las diferencias pre-post del grupo tratado con la media de las diferencias pre-post del grupo control:

$$t = d_{tto} / (1/n_{tto} + 1/n_{ctrl})^{1/2}$$

Este estadístico t elevado al cuadrado coincide con el estadístico F correspondiente al efecto de la interacción en un diseño pre-post con grupo control. Por tanto, teniendo en cuenta que es posible calcular $\hat{\omega}^2$ a partir de F , el valor que cabe esperar que tome $\hat{\omega}^2$ viene dado por:

$$\hat{\omega}^2 = \frac{gl_{interacción} (F_{interacción} - 1)}{gl_{interacción} (F_{interacción} - 1) + N} = \frac{t^2 - 1}{t^2 - 1 + 4n}$$

Así, con $d_{tto} = 1.2$ y $n_{tto} = n_{ctrl} = 25$, se obtiene $t = 1.2 / (1/25 + 1/25)^{1/2} = 4.243$. De ahí se obtiene $\hat{\omega}^2 = (4.243^2 - 1) / (4.243^2 - 1 + 4(25)) = 0.145$, que es, según puede apreciarse en Tabla 1, el valor aproximado que toma $\hat{\omega}^2$ en las diferentes condiciones simuladas.

Por lo que respecta al estadístico d_{PPC} , su valor esperado δ_{PPC} también viene dado por el valor elegido para δ_{tto} . Asumiendo que la media de las diferencias en el grupo control vale cero,

$$\delta_{PPC} = (\delta_{tto} \times \sigma_{dif}) / \sigma_{pre}$$

Por ejemplo, con $\delta_{tto} = 1.2$ y $\sigma_{dif} = 0.775$ (este es el valor de σ_{dif} que corresponde a $\rho_{XY} = 0.7$), se obtiene $\delta_{PPC} = (1.2 \times 0.775) / 1 = 0.93$, que es, según puede apreciarse

en la Tabla 1, el valor aproximado que toma d_{PPC} en las diferentes condiciones simuladas.

En la Tabla 2 se informa de los resultados relativos a uno de los índices de cambio, la diferencia individual tipificada (SID). Los valores de cada casilla se han obtenido promediando los n valores SID de cada muestra (un valor SID por cada caso simulado) y luego calculando la media y la desviación típica de esas 500 medias. La Tabla 2 incluye información de los grupos control y tratado en las 21 condiciones con $n = 25$ y $\rho_{XY} = 0.7$. Los resultados relativos a RCI y al resto de condiciones de la simulación son muy similares a los que incluye la Tabla 2 y no proporcionan información adicional de interés. Todos ellos pueden consultarse en el enlace de material complementario.

Dado que SID se calcula dividiendo cada diferencia pre-post entre la desviación típica de las diferencias, es esperable que la media de SID en cada condición simulada coincida con el tamaño del efecto impuesto en esa condición y grupo (δ_{tto} o δ_{ctrl}). Los resultados de la Tabla 2 parecen indicar que esto es, de hecho, lo que ocurre, *independientemente de la distribución de las puntuaciones*. Además, conforme aumenta el tamaño muestral, esta correspondencia aumenta también al tiempo que las desviaciones típicas disminuyen.

Tabla 2.

Media (desv. típica) del índice *SID* para los grupos control (*ctrl*) y experimental (*tto*) en todas las distribuciones simuladas con $n = 25$ y $\rho_{XY} = 0.7$

δ_{tto}	Grupo	$As = -3$ $Cr = 18$	$As = -2$ $Cr = 9$	$As = -1$ $Cr = 2$	$As = 0$ $Cr = 0$	$As = 1$ $Cr = 2$	$As = 2$ $Cr = 9$	$As = 3$ $Cr = 18$
.0	<i>ctrl</i>	.04 (.23)	.04 (.22)	.02 (.21)	.00 (.20)	-.01 (.20)	-.04 (.22)	-.07 (.22)
	<i>tto</i>	.04 (.23)	.03 (.23)	.00 (.21)	.01 (.21)	-.01 (.21)	-.04 (.22)	-.03 (.22)
.3	<i>ctrl</i>	.04 (.23)	.04 (.23)	.01 (.21)	.01 (.21)	-.03 (.20)	-.02 (.22)	-.04 (.23)
	<i>tto</i>	.34 (.21)	.33 (.22)	.30 (.21)	.29 (.21)	.30 (.21)	.28 (.21)	.27 (.21)
.6	<i>ctrl</i>	.04 (.23)	.02 (.21)	.02 (.22)	.00 (.21)	-.02 (.21)	-.04 (.22)	-.04 (.23)
	<i>tto</i>	.62 (.20)	.62 (.21)	.61 (.21)	.60 (.21)	.56 (.22)	.57 (.21)	.55 (.24)
.9	<i>ctrl</i>	.04 (.23)	.03 (.21)	.02 (.21)	-.01 (.21)	-.01 (.20)	-.02 (.21)	-.04 (.23)
	<i>tto</i>	.94 (.22)	.93 (.22)	.93 (.20)	.90 (.21)	.88 (.21)	.87 (.22)	.87 (.22)
1.2	<i>ctrl</i>	.05 (.24)	.02 (.21)	.01 (.22)	-.02 (.21)	-.03 (.22)	-.02 (.22)	-.06 (.22)
	<i>tto</i>	1.23 (.22)	1.22 (.21)	1.23 (.21)	1.19 (.20)	1.19 (.21)	1.17 (.22)	1.18 (.22)
1.5	<i>ctrl</i>	.04 (.23)	.03 (.21)	.01 (.21)	.01 (.22)	-.01 (.22)	-.02 (.22)	-.05 (.23)
	<i>tto</i>	1.55 (.23)	1.53 (.22)	1.53 (.19)	1.49 (.21)	1.49 (.21)	1.48 (.21)	1.47 (.23)
1.8	<i>ctrl</i>	.04 (.23)	.04 (.22)	.02 (.21)	.01 (.20)	-.02 (.22)	-.04 (.23)	-.03 (.24)
	<i>tto</i>	1.83 (.23)	1.84 (.22)	1.80 (.20)	1.81 (.21)	1.77 (.22)	1.78 (.22)	1.76 (.22)
2.1	<i>ctrl</i>	.03 (.22)	.02 (.21)	.00 (.21)	.01 (.21)	-.02 (.21)	-.04 (.22)	-.04 (.23)
	<i>tto</i>	2.14 (.24)	2.13 (.21)	2.11 (.21)	2.09 (.20)	2.10 (.21)	2.07 (.22)	2.05 (.23)
2.4	<i>ctrl</i>	.03 (.21)	.03 (.22)	.01 (.20)	-.02 (.21)	-.01 (.21)	-.03 (.22)	-.05 (.22)
	<i>tto</i>	2.43 (.22)	2.43 (.22)	2.41 (.22)	2.38 (.21)	2.37 (.22)	2.37 (.23)	2.36 (.22)
2.7	<i>ctrl</i>	.06 (.24)	.03 (.22)	.03 (.21)	.00 (.21)	-.02 (.21)	-.03 (.23)	-.05 (.21)
	<i>tto</i>	2.74 (.23)	2.72 (.22)	2.72 (.22)	2.69 (.21)	2.69 (.21)	2.68 (.21)	2.65 (.22)
3.0	<i>ctrl</i>	.03 (.22)	.01 (.22)	.04 (.21)	-.01 (.21)	.00 (.21)	-.02 (.21)	-.03 (.22)
	<i>tto</i>	3.04 (.24)	3.03 (.21)	3.03 (.22)	3.00 (.21)	2.98 (.21)	2.98 (.21)	2.96 (.22)
3.3	<i>ctrl</i>	.03 (.21)	.03 (.22)	.02 (.22)	-.02 (.21)	-.02 (.21)	-.04 (.23)	-.04 (.22)
	<i>tto</i>	3.34 (.23)	3.32 (.21)	3.32 (.22)	3.30 (.21)	3.28 (.22)	3.28 (.23)	3.26 (.22)
3.6	<i>ctrl</i>	.05 (.22)	.03 (.22)	.02 (.21)	.00 (.21)	-.02 (.23)	-.04 (.21)	-.05 (.23)
	<i>tto</i>	3.64 (.22)	3.62 (.21)	3.62 (.23)	3.61 (.21)	3.60 (.21)	3.55 (.23)	3.56 (.24)

Finalmente, en la Tabla 3 se informa de las medidas del tamaño del efecto elegidas para cuantificar el cambio basado en los centros (con $n = 25$ y $\rho_{XY} = 0.7$). La tabla incluye la media (desviación típica) de los estadísticos $\hat{\omega}^2$ y d_{PPC} , junto con el coeficiente de correlación de Pearson entre ellos. Los valores empíricos de ambos estadísticos se aproximan sus valores esperables a partir del valor de δ_{tto} y δ_{ctrl} y las ecuaciones expuestas anteriormente, lo que indica que la simulación ha

funcionado correctamente. La correlación entre ambos estadísticos es lo bastante alta como para asumir que ambos estadísticos ofrecen información muy similar (sus valores oscilan entre 0.85 y 0.96).

Tabla 3.

Media (dev. típica) de $\hat{\omega}^2$ y d_{PPC} , y coeficiente de correlación de Pearson entre ellos en todas las condiciones simuladas con $n = 25$ y $\rho_{XY} = 0.7$

medida de tamaño del efecto	δ_{tto}	$As = -3$ $Cr = 18$	$As = -2$ $Cr = 9$	$As = -1$ $Cr = 2$	$As = 0$ $Cr = 0$	$As = 1$ $Cr = 2$	$As = 2$ $Cr = 9$	$As = 3$ $Cr = 18$
$\hat{\omega}^2$.0	.01 (.01)	.01 (.01)	.00 (.01)	.00 (.01)	.00 (.01)	.01 (.01)	.00 (.01)
	.3	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)	.02 (.02)	.01 (.02)	.02 (.03)
	.6	.04 (.03)	.04 (.03)	.04 (.04)	.04 (.04)	.04 (.04)	.05 (.04)	.05 (.05)
	.9	.08 (.04)	.08 (.04)	.09 (.05)	.09 (.06)	.09 (.06)	.09 (.06)	.09 (.07)
	1.2	.14 (.05)	.15 (.05)	.15 (.06)	.15 (.06)	.15 (.08)	.15 (.08)	.16 (.09)
	1.5	.21 (.06)	.21 (.06)	.22 (.06)	.21 (.08)	.21 (.09)	.21 (.10)	.22 (.11)
	1.8	.28 (.07)	.28 (.07)	.28 (.07)	.28 (.08)	.28 (.10)	.28 (.11)	.27 (.12)
	2.1	.34 (.07)	.34 (.06)	.35 (.07)	.34 (.08)	.35 (.10)	.34 (.11)	.33 (.13)
	2.4	.41 (.08)	.40 (.07)	.41 (.06)	.40 (.08)	.40 (.10)	.40 (.12)	.40 (.14)
	2.7	.46 (.08)	.46 (.08)	.46 (.07)	.46 (.07)	.47 (.10)	.47 (.12)	.45 (.14)
	3.0	.51 (.09)	.51 (.07)	.52 (.07)	.52 (.07)	.51 (.10)	.51 (.11)	.50 (.14)
	3.3	.56 (.08)	.56 (.07)	.57 (.06)	.57 (.07)	.56 (.09)	.56 (.12)	.54 (.13)
	3.6	.59 (.08)	.60 (.07)	.61 (.06)	.61 (.06)	.61 (.09)	.59 (.11)	.59 (.13)
d_{PPC}	.0	.00 (.24)	-.01 (.23)	-.01 (.23)	.00 (.22)	.00 (.22)	.00 (.24)	.02 (.24)
	.3	.24 (.21)	.24 (.23)	.23 (.22)	.22 (.22)	.26 (.25)	.23 (.26)	.25 (.29)
	.6	.50 (.24)	.48 (.23)	.46 (.22)	.47 (.24)	.46 (.26)	.48 (.31)	.47 (.35)
	.9	.73 (.29)	.70 (.23)	.71 (.24)	.72 (.26)	.68 (.29)	.68 (.34)	.73 (.42)
	1.2	.98 (.38)	.96 (.29)	.95 (.27)	.94 (.27)	.95 (.34)	.95 (.43)	.99 (.50)
	1.5	1.17 (.41)	1.19 (.36)	1.20 (.31)	1.16 (.31)	1.18 (.39)	1.19 (.49)	1.22 (.59)
	1.8	1.48 (.54)	1.42 (.44)	1.40 (.34)	1.40 (.32)	1.41 (.44)	1.42 (.52)	1.40 (.63)
	2.1	1.73 (.62)	1.64 (.42)	1.65 (.38)	1.62 (.35)	1.68 (.48)	1.65 (.61)	1.67 (.74)
	2.4	1.98 (.71)	1.89 (.56)	1.86 (.42)	1.86 (.40)	1.84 (.48)	1.95 (.75)	1.97 (1.09)
	2.7	2.12 (.77)	2.12 (.63)	2.08 (.44)	2.08 (.40)	2.14 (.56)	2.19 (.73)	2.18 (.99)
	3.0	2.46 (.89)	2.40 (.67)	2.35 (.50)	2.33 (.45)	2.29 (.55)	2.36 (.85)	2.39 (1.00)
	3.3	2.68 (1.04)	2.66 (.82)	2.56 (.54)	2.58 (.50)	2.56 (.66)	2.60 (.87)	2.63 (1.17)
	3.6	2.88 (1.01)	2.87 (.84)	2.78 (.57)	2.81 (.52)	2.81 (.72)	2.83 (1.00)	2.89 (1.21)
correlación entre medidas		.847	.899	.943	.961	.944	.906	.858

Después de comprobar la correspondencia entre los datos simulados y los valores impuestos para la simulación, hemos estudiado la relación entre los diferentes métodos utilizados para valorar el cambio: los basados en los centros ($\hat{\omega}^2$ y d_{PPC}) y el basado en los individuos (porcentaje neto de mejorías P_{NM}). Para ello, en primer lugar hemos creado un conjunto de diagramas de dispersión con el estadístico basado en los centros en el eje horizontal y el P_{NM} en el eje vertical. En la Figura 1 se muestran los diagramas de dispersión creados utilizando $\hat{\omega}^2$ y P_{NM} . En la Figura 2 se muestran los diagramas de dispersión creados utilizando d_{PPC} y P_{NM} . En ambos casos los resultados están calculados a partir de *SID*, con $n = 100$ y $\rho_{XY} = 0.7$. Cada uno de los puntos en estos diagramas representan una de las muestras simuladas (que incluye un grupo control y un grupo tratado): los 13 tamaños de efecto multiplicados por las 500 muestras simuladas dieron lugar a 6500 puntos por diagrama. Para facilitar la interpretación de la relación subyacente, mostramos aquí los diagramas correspondientes a las condiciones con mayor tamaño muestral; sin embargo, se observan los mismos patrones con $n = 25$ y $n = 50$, aunque con una mayor variabilidad: para un determinado tamaño de efecto, a menor tamaño muestral, mayor dispersión de los puntos en torno al mismo centro. Con $n = 10$ la variabilidad es tan alta que es difícil interpretar la tendencia. Con $\hat{\omega}^2$, la relación no cambia en los distintos valores de ρ_{XY} . Con d_{PPC} , la forma de la relación se mantiene idéntica al cambiar ρ_{XY} , aunque la pendiente sí cambia (profundizaremos sobre esto más adelante). Los resultados relativos al P_{NM} basado en *RCI* son prácticamente idénticos a los mostrados en las Figuras 1 y 2.

Figura 1

Diagramas de dispersión con el tamaño del efecto basado en los centros ($\hat{\omega}^2$) en el eje horizontal y el porcentaje neto de mejoras (P_{NM} , basado en *SID*) en el eje vertical ($n = 100$, $\rho_{XY} = 0.7$)

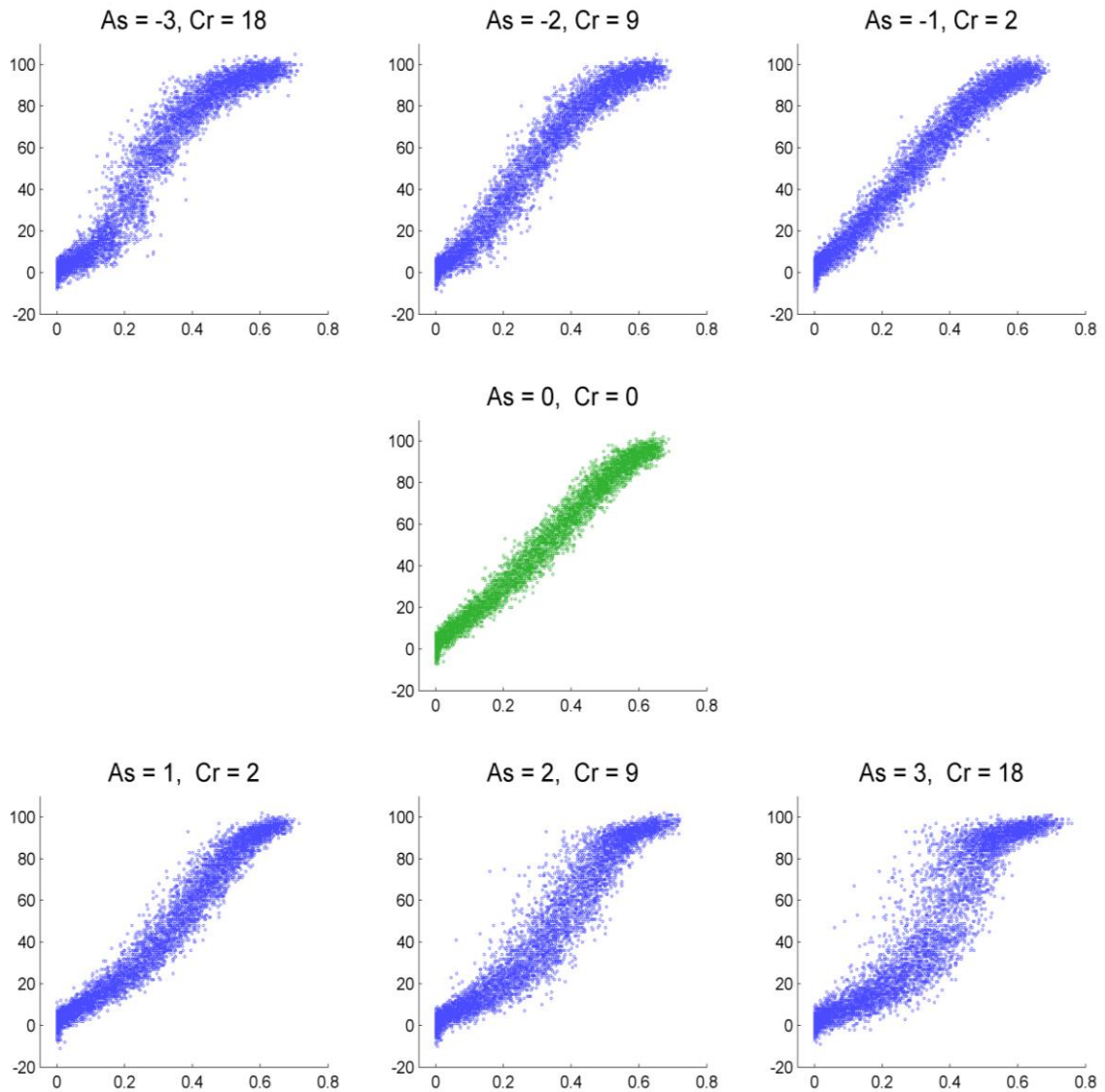
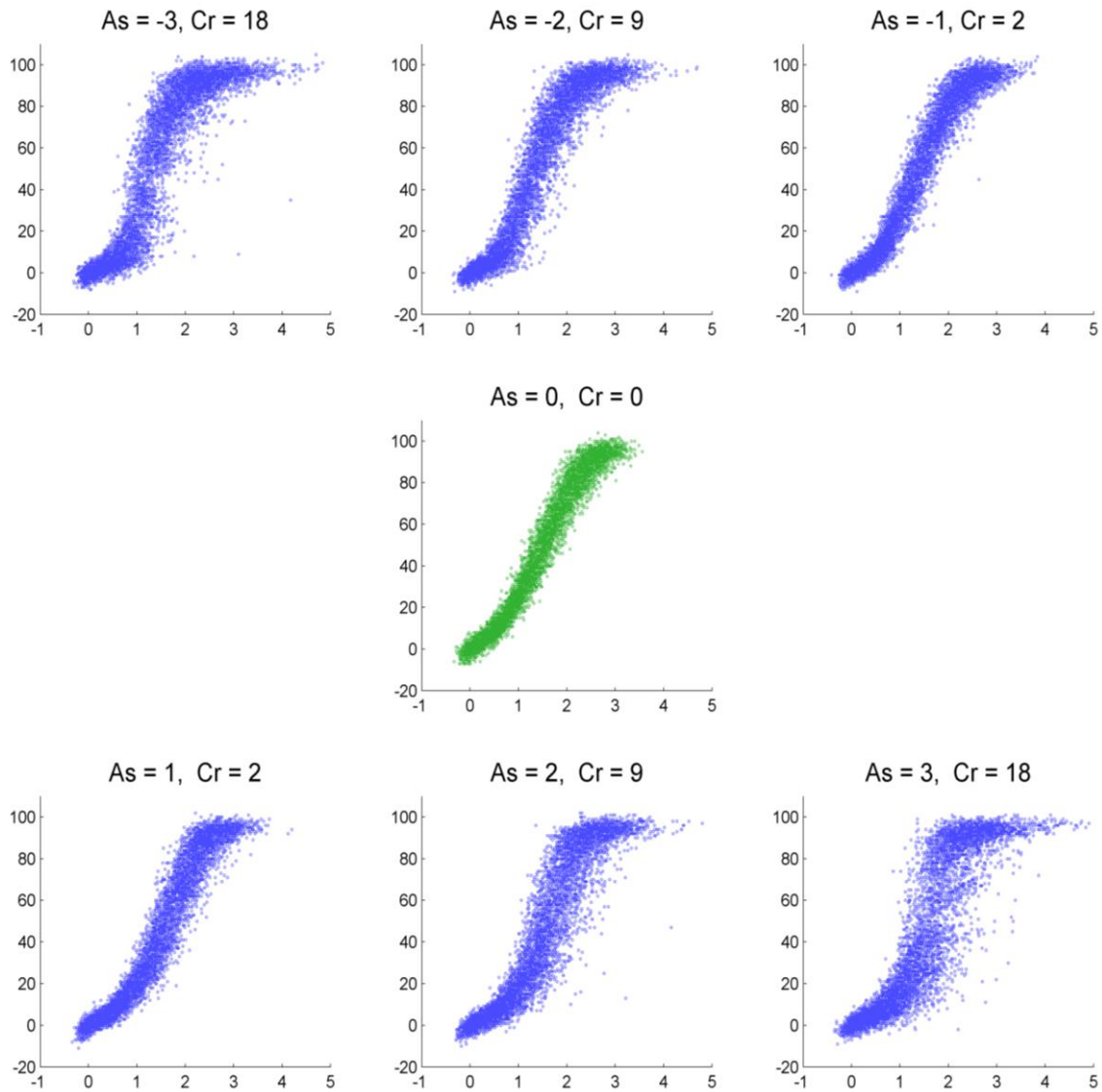


Figura 2

Diagramas de dispersión con el tamaño del efecto basado en los centros (d_{PPC}) en el eje horizontal y el porcentaje neto de mejorías (P_{NM} , basado en SID) en el eje vertical ($n = 100$, $\rho_{XY} = 0.7$)



En los diagramas de dispersión de las Figuras 1 y 2 se aprecia con claridad una relación monótona creciente. Para cuantificar esta relación, hemos estimado cuatro funciones de regresión: lineal, cuadrática, cúbica y logística. En todos los casos hemos utilizado $\hat{\omega}^2$ y d_{PPC} por separado como variables independientes y el P_{NM} como variable dependiente. Las cuatro funciones fueron estimadas para cada una de las 84 condiciones simuladas. En la Tabla 4 informamos del coeficiente de determinación (R^2) para todas las funciones. Los resultados de la Tabla 4 se basan en el P_{NM} calculado con *SID* y $n = 25$. Dado que la dispersión de los diagramas aumenta a medida que el tamaño muestral disminuye, los valores de R^2 de la Tabla 4 son menores que los obtenidos con $n = 50$ y $n = 100$, aunque mayores que los obtenidos con $n=10$. La última fila de la tabla incluye los valores medio, mínimo y máximo asociados a cada función. Con el porcentaje de cambios basado en *RCI* se consigue un ajuste muy similar, aunque ligeramente inferior. Los resultados para el resto de condiciones se pueden consultar en el enlace de material complementario.

Tabla 4

R^2 de las funciones lineal, cuadrática, cúbica y logística para las condiciones con $n = 25$. Variable dependiente: P_{NM} basado en SID

Distribución	$\hat{\omega}^2$ como variable independ.					d_{PPC} como variable independ.			
	ρ_{XY}	Lineal	Cuad.	Cub.	Log.	Lineal	Cuad.	Cub.	Log.
$As = -3$ $Cr = 18$.5	.857	.881	.885	.858	.816	.850	.851	.817
	.7	.857	.887	.890	.858	.834	.883	.883	.836
	.9	.841	.886	.888	.843	.829	.892	.892	.832
$As = -2$ $Cr = 9$.5	.894	.907	.911	.894	.862	.882	.883	.862
	.7	.891	.912	.915	.892	.873	.902	.903	.873
	.9	.887	.917	.920	.888	.879	.918	.920	.880
$As = -1$ $Cr = 2$.5	.916	.924	.926	.916	.891	.902	.903	.891
	.7	.916	.927	.929	.916	.900	.915	.916	.900
	.9	.916	.931	.933	.917	.908	.926	.928	.908
$As = 0$ $Cr = 0$.5	.926	.929	.930	.926	.910	.913	.914	.909
	.7	.928	.930	.931	.927	.914	.917	.918	.913
	.9	.924	.926	.928	.924	.919	.923	.924	.919
$As = 1$ $Cr = 2$.5	.906	.906	.908	.905	.893	.894	.896	.893
	.7	.902	.902	.903	.901	.898	.898	.900	.898
	.9	.898	.898	.900	.898	.896	.896	.898	.896
$As = 2$ $Cr = 9$.5	.866	.867	.868	.867	.864	.867	.869	.864
	.7	.865	.865	.867	.865	.868	.869	.872	.868
	.9	.841	.841	.843	.842	.848	.848	.850	.848
$As = 3$ $Cr = 18$.5	.824	.825	.827	.825	.826	.829	.832	.826
	.7	.815	.816	.817	.816	.826	.827	.829	.826
	.9	.785	.786	.787	.785	.795	.795	.797	.796
Valor medio		.879	.889	.891	.879	.869	.883	.885	.869
Valor mín.		.785	.786	.787	.785	.795	.795	.797	.796
Valor máx.		.928	.931	.933	.917	.919	.923	.924	.919

En general, las cuatro funciones consiguen un ajuste muy bueno. El ajuste es tanto mejor cuanto más se aproximan a la normalidad las distribuciones simuladas. Con las funciones cuadrática y cúbica se obtiene un ajuste ligeramente mejor que con la lineal, aunque con asimetrías no negativas la diferencia es mínima; con la función logística se consigue un ajuste similar al lineal. Puesto que los modelos más parsimoniosos son preferibles a los más complejos, el pequeño incremento en el ajuste que ofrecen los modelos cuadrático y cúbico no justifica su elección frente al

modelo lineal. Por otro lado, el modelo lineal es mucho más sencillo de interpretar que el logístico, especialmente para el profesional e investigador aplicado.

En la Tabla 5 informamos de los coeficientes de la función lineal con $n = 25$. Estos coeficientes permiten estimar el P_{NM} a partir de las medidas de tamaño del efecto basadas en los centros. Los resultados para el resto de condiciones se pueden consultar en el enlace de material complementario.

Tabla 5

Coeficientes (y errores típicos) para el modelo de regresión lineal ($n = 25$, P_{NM} basado en SID)

	$\rho_{XY} = .5$		$\rho_{XY} = .7$		$\rho_{XY} = .9$	
	B_0	B_1	B_0	B_1	B_0	B_1
$\hat{\omega}^2$						
$As = -3, Cr = 18$	3.51 (.29)	162.9 (.82)	4.18 (.29)	164.7 (.83)	5.92 (.31)	164.9 (.89)
$As = -2, Cr = 9$	3.02 (.24)	160.3 (.69)	3.95 (.25)	162.0 (.70)	5.18 (.25)	162.5 (.72)
$As = -1, Cr = 2$	3.49 (.21)	156.7 (.59)	3.74 (.21)	157.5 (.59)	4.50 (.21)	158.2 (.59)
$As = 0, Cr = 0$	2.35 (.19)	152.2 (.53)	2.56 (.19)	151.7 (.53)	2.54 (.19)	151.8 (.54)
$As = 1, Cr = 2$	1.07 (.21)	149.8 (.60)	.86 (.22)	147.6 (.61)	.37 (.22)	146.6 (.61)
$As = 2, Cr = 9$.96 (.26)	148.8 (.73)	.33 (.26)	146.2 (.72)	-.03 (.28)	143.4 (.77)
$As = 3, Cr = 18$.95 (.30)	147.6 (.84)	.63 (.31)	144.1 (.85)	.43 (.33)	139.7 (.91)
d_{PPC}						
$As = -3, Cr = 18$	8.05 (.46)	21.91 (.20)	1.48 (.48)	27.27 (.26)	15.19 (.52)	43.98 (.49)
$As = -2, Cr = 9$	3.47 (.39)	24.29 (.17)	6.17 (.41)	30.42 (.23)	8.51 (.43)	51.35 (.42)
$As = -1, Cr = 2$.04 (.32)	26.46 (.14)	1.18 (.33)	33.81 (.20)	3.12 (.35)	56.61 (.35)
$As = 0, Cr = 0$	-1.64 (.28)	26.38 (.13)	-1.41 (.28)	33.68 (.17)	-1.38 (.29)	58.16 (.29)
$As = 1, Cr = 2$	-1.39 (.32)	24.79 (.15)	-1.32 (.33)	31.32 (.19)	-1.18 (.33)	52.95 (.33)
$As = 2, Cr = 9$	2.34 (.39)	22.00 (.17)	2.46 (.39)	27.48 (.22)	3.19 (.40)	45.02 (.39)
$As = 3, Cr = 18$	6.37 (.45)	19.14 (.19)	6.66 (.45)	23.66 (.24)	6.84 (.45)	38.61 (.41)

Con $\hat{\omega}^2$ como variable independiente, la intersección (B_0) toma valores entre -0.04 y aproximadamente 6, con una media de 2.41 y errores típicos que oscilan entre 0.19 y 0.33). La pendiente (B_1) toma valores entre 140 y 165, con una media de 153 y errores típicos que oscilan entre 0.53 y 0.91. Tomando como ejemplo los resultados obtenidos con distribuciones normales, estos coeficientes indican que

para una intervención con un efecto nulo ($\hat{\omega}^2 = 0$), la función lineal estima un porcentaje neto de cambios individuales de aproximadamente el 2.5%; y por cada 0.10 puntos adicionales en $\hat{\omega}^2$ la función estima que el porcentaje neto de cambios individuales aumenta aproximadamente 15 puntos (obviamente, dado que estamos pronosticando porcentajes, los valores menores de cero y mayores de 100 deben reemplazarse por sus respectivos límites). A la luz de estos resultados, los valores de $\hat{\omega}^2$ propuestos para identificar efectos de tamaño pequeño (0.01), mediano (0.06) y grande (0.14) (Kirk, 2013, p. 138) equivaldrían a un P_{NM} de 4%, 12% y 24% respectivamente. En la Tabla 5 se aprecia que los coeficientes B_0 y B_1 no cambian sustancialmente al cambiar la correlación pre-post. Con el resto de tamaños muestrales se obtienen resultados similares (ver enlace de material complementario).

A diferencia de lo que ocurre con $\hat{\omega}^2$, los coeficientes de regresión correspondientes al estadístico d_{PPC} se ven muy afectados por el grado de relación entre el pre y el post: el valor de los coeficientes va aumentando conforme lo hace la relación pre-post (este efecto es especialmente notable en la pendiente B_1). Tomando como referencia las condiciones con $\rho_{XY} = 0.7$, el coeficiente para la intersección B_0 está entre -1.4 y aproximadamente 10, con una media de 3.5 y errores típicos que oscilan entre 0.19 y 0.28. La pendiente (B_1) toma valores entre 23.7 y 33.8, con una media de 30 y errores típicos que oscilan entre 0.17 y 0.27. Con distribuciones normales, estos coeficientes indican que, cuando el efecto de la intervención es nulo ($d_{PPC} = 0$), cabe esperar un porcentaje neto de cambios de -1.4% (es decir, de 0%); y, por cada punto adicional en d_{PPC} , cabe esperar que el porcentaje neto de cambios aumente 33.68 puntos. Los valores de d_{PPC} propuestos por Cohen (1969) para definir un efectos de tamaño pequeño, mediano y grande (0.2, 0.5 y 0.8) equivaldrían a 5%, 15% y 26% de P_{NM} , respectivamente.

Es importante señalar que la relación entre d_{PPC} y P_{NM} depende de la intensidad de la relación entre las puntuaciones pre y post. Para ilustrar esta idea,

en la Tabla 6 se muestra la correspondencia entre estos dos estadísticos con distintos valores de ρ_{XY} . En consecuencia, para poder usar d_{PPC} como estimador del *porcentaje neto de mejorías* es necesario tener en cuenta la relación entre las puntuaciones pre y post.

Tabla 6

Relación entre d_{PPC} y P_{NM} basado en *SID* con $n = 25$ y distribuciones normales.

	$\rho_{XY} = .5$	$\rho_{XY} = .7$	$\rho_{XY} = .9$
$d_{PPC} = .2$ implica un % de cambios de	3.6%	5.3%	1.3%
$d_{PPC} = .5$ implica un % de cambios de	11.6%	15.4%	27.7%
$d_{PPC} = .8$ implica un % de cambios de	19.5%	25.5%	45.1%
100% de cambios se alcanza con $d_{PPC} =$	3.73	2.93	1.70

Discusión

El principal objetivo de este estudio era comprobar si el cambio en los centros de las distribuciones (cuantificado mediante medidas convencionales del tamaño del efecto) está relacionado con el cambio basado en los individuos (cuantificado mediante el porcentaje neto de mejorías, P_{NM}), cuando ambos se obtienen comparando el cambio observado en un grupo tratado con el cambio observado en un grupo control. Los diagramas de dispersión de la Figura 1 muestran que el porcentaje neto de mejorías no es independiente del tamaño del efecto: a medida que el tamaño del efecto aumenta, también lo hace el porcentaje de cambios.

Nuestro segundo objetivo era encontrar una función matemática capaz de dar cuanta de esa relación. Y creemos haberla encontrado. Aunque con todos los modelos de regresión que hemos ajustado hemos obtenidos buenos valores de ajuste, el modelo lineal es el más parsimonioso y fácil de interpretar. Por ello, consideramos que el modelo lineal es preferible al cuadrático, cúbico o logístico. El

grado de ajuste obtenido con el modelo lineal está entre muy bueno y excelente (con $n = 25$ se obtienen valores de R^2 que oscilan entre 0.79 y 0.93; ver Tabla 4).

Junto con lo anterior, el hallazgo más interesante de este tercer estudio es que, cuando usamos el estadístico $\hat{\omega}^2$ como medida del tamaño del efecto (y, por tanto, como variable independiente para pronosticar el porcentaje neto de mejorías), *la pendiente del modelo de regresión lineal es aproximadamente la misma en todas las condiciones simuladas*: alrededor de 153 puntos (dependiendo de las condiciones simuladas, toma valores que oscilan entre 140 y 165). Puesto que los valores de $\hat{\omega}^2$ van de 0 a 1, quizá sería más útil y apropiado plantear esta conclusión en los siguientes términos: por cada 0.10 puntos que aumenta el tamaño del efecto (cuantificado con $\hat{\omega}^2$), la función lineal estima que el porcentaje de cambios aumenta en 15.3 puntos (entre 14 y 16.5 puntos dependiendo de las condiciones particulares de las distribuciones simuladas).

Los resultados encontrados con $\hat{\omega}^2$ permiten concluir que, bajo determinadas circunstancias, cuando el porcentaje neto de mejorías es desconocido, es posible estimarlo basándonos en el tamaño del efecto grupal, que habitualmente es conocido.

A diferencia de lo que ocurre con $\hat{\omega}^2$, los coeficientes de la regresión lineal para el estadístico d_{PPC} se ven fuertemente afectados por la intensidad de la relación entre las puntuaciones pre y post. Sin embargo, esto es exactamente lo que cabe esperar si tenemos en cuenta que el estadístico d_{PPC} se basa en la desviación típica de las puntuaciones pre (que no tiene en cuenta el grado de relación entre el pre y el post), mientras que P_{NM} se obtiene aplicando índices de cambio (SID , RCI) basados en las desviaciones típicas de las diferencias pre-post (que sí tienen en cuenta el grado de relación pre-post). Sin embargo, si se conoce la desviación típica de las diferencias pre-post y la desviación típica en pre, es posible transformar d_{PPC} en $\hat{\omega}^2$ para conseguir una estimación más precisa del porcentaje neto de mejorías. Para ello se debe aplicar la expresión:

$$\hat{\omega}^2 = \frac{d_{PPC}^2 n_{tto} n_{ctrl} S_{pre}^2 - (n_{tto} + n_{ctrl}) S_{dif}^2}{d_{PPC}^2 n_{tto} n_{ctrl} S_{pre}^2 + (n_{tto} + n_{ctrl})(4n_{tto} - 1) S_{dif}^2}$$

Nótese que, por un lado, d_{PPC} es función de d_{tto} , d_{ctrl} , S_{dif} y S_{pre} , mientras que por otro lado, S_{dif} es función de S_{pre} , S_{post} y la covarianza pre-post (ver el comentario a la Tabla 1 de este estudio). Por lo tanto, esta expresión puede modificarse de acuerdo con la información disponible en cada estudio.

Nuestros resultados nos permiten afirmar que la relación encontrada entre los estadísticos que evalúan el cambio a partir de los centros de las distribuciones y los estadísticos que evalúan el cambio a partir de las puntuaciones individuales (en diseños pre-post con grupos aleatorios) se encuentra en todas las condiciones que hemos simulado. No obstante, el ajuste de la función de regresión lineal varía entre ellas: en las condiciones más favorables (distribuciones pre y post normales), R^2 alcanza un valor de 0.93; en las más desfavorables (asimetría positiva extrema), baja hasta 0.79. A medida que el tamaño muestral aumenta (ver enlace de material complementario), también lo hace el ajuste: con $n = 100$, por ejemplo, R^2 alcanza valores de 0.98 en las condiciones más favorables, y nunca baja de 0.87 en las más desfavorables.

Como conclusión general de este estudio, podemos afirmar que, en los diseños pre-post con grupos aleatorios el cambio basado en los centros se puede considerar linealmente relacionado con el cambio basado en los individuos. En consecuencia, es posible, bajo ciertas circunstancias, estimar el porcentaje neto aproximado de mejorías debidas a un tratamiento en un contexto aplicado incluso cuando solo se conocen estadísticos basados en los centros. La relación que hemos encontrado se puede resumir en la idea de que por cada 0.10 puntos de incremento en $\hat{\omega}^2$ cabe esperar que el porcentaje de cambios individuales aumente un 15%. Con ligeras variaciones, este patrón se observa en todas las condiciones simuladas.

Discusión general

El objetivo de este trabajo era estudiar la posible relación entre dos formas distintas de valorar el cambio tras una intervención. La primera de estas formas es la que hemos denominado *cambio basado en los centros o promedios*, y consiste en decidir si ha habido un cambio comparando (básicamente) los centros de las distribuciones antes y después de la intervención. La segunda es la que hemos denominado *cambio basado en los individuos*, y consiste en aplicar índices de cambio individual para decidir qué individuos han cambiado, y posteriormente calcular el porcentaje de cambios.

Nuestros resultados nos permiten afirmar que estos dos enfoques no solo no son independientes entre sí, sino que están muy relacionados:

1. Los resultados del primer estudio muestran que en un *diseño pre-post de grupo único* (un solo grupo medido en dos momentos, antes y después de la intervención) existe una relación estrecha entre la medida del tamaño del efecto basada en los centros ($d = M_{\text{dif}} / S_{\text{dif}}$) y el porcentaje de cambios (casos individuales con $SID > 1.64$). Esta relación se resume en la idea de que, independientemente de la forma de las distribuciones, el porcentaje de cambios aumenta aproximadamente en 30 puntos por cada punto que aumenta d . Esto implica que la relación entre ambas formas de medir el cambio se puede

considerar lineal. Con tamaños muestrales a partir de 25 casos, la función lineal nunca arroja valores de ajuste (R^2) por debajo de 0.90.

2. Los resultados del segundo estudio muestran que, en un diseño de *grupos aleatorios* (dos grupos, uno tratado y otro control, que son medidos solo después de la intervención), también existe una relación fuerte entre el tamaño del efecto basado en los centros (d y R_{bp}) y el cambio basado en los individuos (*probabilidad de superioridad* de una puntuación del grupo tratado respecto a una puntuación del grupo control; PS). La relación entre d y PS se puede modelar mediante una regresión cuadrática, independientemente de la forma de las distribuciones. Con tamaños muestrales a partir de 25 casos por grupo, esta función arroja valores de ajuste (R^2) por encima de 0.89. La relación entre R_{bp} y PS se puede modelar linealmente, también con independencia de la forma de las distribuciones. Con tamaños muestrales a partir de 25 casos por grupo, esta función arroja valores de ajuste (R^2) por encima de 0.90.

3. Los resultados del tercer estudio muestran que, en un diseño *pre-post con grupos aleatorios* (dos grupos, uno tratado y otro control, ambos medidos antes y después de la intervención), el cambio en los centros, medido mediante $\hat{\omega}^2$ o mediante d_{PPC} , está estrechamente relacionado con el *porcentaje neto de mejoras individuales* (P_{NM}). Tanto si se utiliza $\hat{\omega}^2$ como si se utiliza d_{PPC} para medir el tamaño del efecto, la relación encontrada entre ellos y el P_{NM} es de tipo lineal. El grado de ajuste que se obtiene utilizando estos estadísticos como predictores del P_{NM} es entre muy bueno y excelente; a partir de 25 casos por grupo, los valores de R^2 oscilan entre 0.79 y 0.93. Por cada 0.10 puntos de incremento en $\hat{\omega}^2$ el valor del P_{NM} aumenta aproximadamente 15.3 puntos, con valores que oscilan entre 14 y 16.5 dependiendo de la forma de la distribución.

Cuando se utiliza d_{PPC} , el valor de la pendiente depende de la correlación pre-post R_{XY} , puesto que esta afecta a la métrica de d_{PPC} .

Al margen de los valores específicos de cada contexto, lo realmente interesante de la relación detectada es que la información que se obtiene con cada enfoque ayuda a completar y entender mejor la información que se obtiene con el otro enfoque. Lejos de ser independientes, como algunos autores mantienen (ver la introducción), nuestros resultados nos llevan a pensar que ambos enfoques son expresiones distintas de prácticamente la misma información. Por lo tanto, el efecto de cualquier intervención (y, en general, *cualquier efecto* que se desee estudiar) puede ser evaluado mediante ambos enfoques para obtener información equivalente.

Los métodos basados en los individuos tienen la ventaja de ser más fáciles de interpretar, en especial para profesionales e investigadores aplicados. Por ejemplo, informar de que, tras una determinada intervención, el 60% de los individuos han experimentado un cambio positivo es mucho más fácil de entender y admite una interpretación más directa que informar de un tamaño del efecto de $d = 2$ (la media post se aleja dos desviaciones típicas de la media pre). Los dos estadísticos están informando de un efecto de tamaño similar, pero el primer enfoque permite una comunicación de los resultados mucho más sencilla. Esto es aún más acusado con estimaciones del efecto basadas en los centros menos conocidas que d , tales como $\hat{\omega}^2$ y η^2 parcial. Estos estadísticos son relativamente frecuentes en la literatura sobre Análisis de Varianza, y por lo tanto en diseños con dos o más factores. En nuestra opinión, son los mejores estadísticos para valorar el efecto de una intervención en un diseño puramente experimental, que es el que permite la atribución causal del efecto al tratamiento. Sin embargo, son poco conocidos por la mayoría de profesionales e investigadores aplicados. En este contexto, el uso del *porcentaje neto de mejoras* (porcentaje de casos mejoran en el

grupo tratado, descontando el porcentaje de mejorías del grupo control) permite valorar el efecto del tratamiento con la misma precisión pero con la ventaja añadida de ser más fácil de comprender e interpretar.

Por otro lado, los métodos basados en los centros son ampliamente conocidos y utilizados en el ámbito metodológico. Prácticamente cualquier estudio en cualquier rama de la psicología incluye una valoración de los efectos estudiados mediante una estimación basada en los centros de las distribuciones. En el campo del meta-análisis, preocupado precisamente por recopilar y comparar todos los efectos encontrados en distintos estudios primarios sobre un determinado problema, existe abundante literatura dedicada a la evaluación del tamaño del efecto a través de métodos basados en los centros. El presente trabajo permite establecer una conexión entre esos métodos y los basados en los individuos, de forma que sea posible trasladar cualquier efecto del que se informe en un estudio ya publicado a una estimación del porcentaje de cambios, de la probabilidad de superioridad o del porcentaje neto de mejorías. Para facilitar esta traslación, hemos incluido en el Apéndice 1 de esta tesis varias tablas de equivalencias que permiten buscar, para cada diseño, el valor del estadístico basado en los individuos a partir del estadístico basado en los centros.

Uno de los objetivos que persigue este trabajo y, de forma general, nuestra línea de investigación, es encontrar criterios objetivos para evaluar la eficacia de una intervención. En las ciencias sociales y de la salud es habitual la aplicación (muchas veces “a ciegas”) de los criterios propuestos por Cohen (1969, pp. 24-27, 1992) para declarar un efecto como pequeño ($d = 0.2$), mediano ($d = 0.5$) y grande ($d = 0.8$). Como ya hemos señalado en nuestros tres estudios, los resultados obtenidos nos llevan a cuestionar seriamente que estos criterios sean apropiados. Esta postura es coherente con la información que se desprende de estudios previos que han puesto de manifiesto que estos puntos de corte tampoco son apropiados desde el punto de vista de la tasa de falsos negativos que generan (Ferrer, 2014).

Además, la propuesta de Cohen no es la única existente. Como ya hemos señalado, Ferguson (2009) ha propuesto criterios alternativos para declarar un efecto como mínimo ($d = 0.41$), moderado ($d = 1.15$) y fuerte ($d = 2.70$). Desde el punto de vista del cambio basado en los individuos, nuestros resultados indican que estos puntos de corte son más razonables que los de Cohen.

No obstante, uno de los principales hallazgos de este trabajo es que, si se conoce el valor de la estimación del tamaño del efecto basado en los centros, es posible estimar el porcentaje de cambios (o la probabilidad de superioridad, o el porcentaje neto de mejoras) asociado a una determinada intervención (ver Apéndice 1).

Teniendo esto en mente, proponemos que, para valorar la relevancia de un efecto, puede prescindirse por completo de puntos de corte y poner el foco en informar tanto del cambio basado en los centros (tamaño del efecto) como del cambio basado en los individuos (porcentaje de cambios). De este modo, para valorar el efecto de una intervención, no sería necesario comprometerse con puntos de corte arbitrarios, sino que bastaría con presentar la información basada en los centros acompañada de la información basada en los individuos.

Unos ejemplos. En un diseño *pre-post con grupo único*, el efecto del tratamiento se podría informar mediante $d = 1.6$ y *porcentaje de cambios* = 50%. En un diseño de *grupos aleatorios* se podría informar mediante $d = 0.5$ y *probabilidad de superioridad* = 0.64. Y en un diseño *pre-post con grupos aleatorios* se podría informar mediante $\hat{\omega}^2 = 0.15$ y *porcentaje neto de mejoras* = 23%; o bien, mediante $d_{PPC} = 0.94$ y *porcentaje neto de mejoras* = 23% (con $R_{XY} = 0.7$). Procediendo de esta manera no sería necesario declarar que el efecto encontrado es “pequeño”, “mediano” o “grande” para ayudar al lector a interpretar la significación sustantiva de d ; el propio lector la puede deducir del estimador basado en los individuos, el

cual, además, a diferencia del basado en los centros, tiene valores mínimo y máximo.

General Discussion

The objective of this work was to study the possible relationship between two different ways of assessing the change after an intervention. The first of these forms is what we call *centers-based change*, and involves deciding whether there has been a change comparing (basically) the centers of the distributions before and after the intervention. The second is what we call *individuals-based change*, and involves applying individual change indices to decide which individuals changed, and then calculate the percentage changes.

Our results allow us to affirm that these two approaches are not independent, but closely related:

1. The results of the first study show that in a *single group pre-post* design (one group measured in two time points, before and after the intervention) there is a close relationship between the centers-base measure of effect size ($d = M_{\text{dif}} / S_{\text{dif}}$) and percentage of changes (individual cases with *standardized individual difference* $SID > 1.64$). This relationship is summarized in the idea that, regardless of the form of distributions, the percentage of changes raises approximately by 30 points for each additional point of the Cohen's standardized difference d . This entails that the relationship between the two

ways of measuring change can be considered as linear. With sample sizes from 25 cases, the linear function never yields fit values (R^2) below 0.90.

2. The results of the second study show that, in a *randomized groups* design (two groups, one treated and one control, which are measured only after the intervention), there is also a strong relationship between the centers-based effect size measure (Cohen's d and *point-biserial correlation* R_{pb}) and individuals-based change (*probability of superiority* of a score from the treated group compared with a score from the control group; PS). The relationship between d and PS can be modeled by a quadratic regression, regardless of the form of distributions. With sample sizes from 25 cases per group, this function yields fit values (R^2) above 0.89. The relationship between R_{pb} and PS can be modeled as linear, also regardless of the shape of the distributions. With sample sizes from 25 cases per group, this function yields fit values (R^2) above 0.90.
3. The results of the third study show that, in a *randomized groups pre-post* design (two groups, one treated and another control, both measured before and after the intervention) the centers-based change, measured through $\hat{\omega}^2$ or through d_{PPC} , is closely related to the *net percentage of individual improvements* (P_{NM}). Regardless of whether the group change is measured through $\hat{\omega}^2$ or through d_{PPC} , the relationship between them and the P_{NM} is linear. The functions' fit obtained when using these statistics as predictors of P_{NM} ranges between very good and excellent; from 25 cases per group, R^2 values range between 0.79 and 0.93. For each 0.10 point increase in $\hat{\omega}^2$, the value of the P_{NM} increases about 15.3 points, with values ranging between 14 and 16.5 depending on the shape of the distribution. When d_{PPC} is used, the value of the slope depends on the pre-post correlation R_{XY} , since it affects the metric of d_{PPC} .

Aside from the specific values for each context, the most interesting finding about the detected relationship is the fact that the information yielded by each approach helps completing and better understanding the information yielded by the other approach. Far from being independent, as some authors claim (see our introduction), our results lead us to think that both approaches are different expressions of essentially the same information. Therefore, the effect of any intervention (and, in general, *any effect*) can be assessed through both approaches, and equivalent information will be obtained.

The individuals-based methods have the advantage of being easier to interpret, especially for applied practitioners and researchers. For example, reporting that, following an intervention, 60% of individuals experienced a positive change is much easier to understand and allows a more direct interpretation than reporting an effect size of $d = 2$ (the post mean is two standard deviations away from the pre mean). Both statistics report an effect of similar size, but the first approach allows a much easier communication of the results. This is even more pronounced with lesser known centers-based effect estimates, such as $\hat{\omega}^2$ and partial η^2 . These statistics are relatively common in the literature on Analysis of Variance, and therefore in designs with two or more factors. In our opinion, they are the best statistics to assess the effect of an intervention on a purely experimental design, which is what allows the causal attribution of the effect to the treatment. However, they are unfamiliar to most practitioners and applied researchers. In this context, the use of *net percentage of individual improvements* (percentage of improved cases in the treated group, minus the percentage of improved cases in the control group) allows assessing the treatment effect with the same accuracy but with the added advantage of being easier to understand and interpret.

On the other hand, center-based methods are widely known and used in the methodology field. Almost every study from any branch of psychology includes an assessment of the studied effects through an estimation based in the centers of the distributions. In the field of meta-analysis, especially focused in collecting and comparing all the available effects from different primary studies on a particular topic, there is abundant literature on the assessment of the effect size through center-based methods. Our work allows establishing a connection between these methods and individuals-based ones, so that it is possible to transfer any effect reported in an already published study to an estimation of the percentage of changes, the probability of superiority or the net percentage of individual improvements. To facilitate this transfer, we included in Appendix 1 of this thesis several equivalence tables to search, for each design, the value for the individuals-based statistic based on the value of the centers-based statistic.

One of the objectives of this work and, in general, our research line, is to find objective criteria for assessing the effectiveness of an intervention. In the social and health sciences, it is common the (often "blind") application of the criteria proposed by Cohen (1969, pp. 24-27, 1992) to declare an effect as small ($d = 0.2$), medium ($d = 0.5$) and large ($d = 0.8$). As noted in our three studies, the results lead us to seriously question the appropriateness of these criteria. This is consistent with the findings from previous works that showed the inadequateness of these cutoffs also from the point of view of the false negatives and false positives rates (Ferrer, 2014).

Furthermore, Cohen's proposal is not the only option. As noted above, Ferguson (2009) has proposed alternative criteria to declare an effect as minimum ($d = 0.41$), moderate ($d = 1.15$) and strong ($d = 2.70$). From the point of view of change based on the individuals, our results indicate that these cutoffs are more tenable than Cohen's.

However, one of the main findings of the present work is the idea that, if the value of the centers-based estimation for the effect size is known, it is possible to estimate the percentage of changes (or the probability of superiority, or the net percentage of individual improvements) associated to a given intervention (see Appendix 1).

With this in mind, we propose to assess the relevance of an effect by entirely disregarding any cutoff and, instead, reporting both the centers-based change (effect size) and the individuals-based change (percentage of changes). Thus, the researchers do not need to commit to any arbitrary cutoff to assess the effect of an intervention but they just need to accompany the centers-based information with individuals-based information.

Some examples: In a *single group pre-post* design, the treatment effect could be reported by $d = 1.6$ and *percentage of changes* = 50%. In a *randomized groups* design could be reported by $d = 0.5$ and *probability of superiority* = 0.64. And in a *randomized groups pre-post* design it could be reported by $\hat{\omega}^2 = 0.15$ and *net percentage of individual improvements* = 23%; or by $d_{PPC} = 0.94$ and *net percentage of individual improvements* = 23% (with $R_{XY} = 0.7$). By proceeding in this way, it would not be necessary to declare that the effect found is "small", "medium" or "large" to help the reader interpret the substantive significance of d ; the reader himself/herself can deduce it from the individuals-based estimate, which, besides, has minimum and maximum values.

Referencias bibliográficas

- Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, 15(2), 94-98. <http://doi.org/10.1111/j.0963-7214.2006.00414.x>
- Aharony, L., & Strasser, S. (1993). Patient satisfaction: What we know about and what we still need to explore. *Medical Care Research and Review*, 50(1), 49-79. <http://doi.org/10.1177/002570879305000104>
- Albrecht, G., & Hoogstraten, J. (1998). Satisfaction as a determinant of compliance. *Community Dentistry and Oral Epidemiology*, 26(2), 139-146. <http://doi.org/10.1111/j.1600-0528.1998.tb01940.x>
- Ato, M., Lopez-Garcia, J. J., & Benavente, A. (2013). A classification system for research designs in psychology. *Anales de Psicología*, 29(3), 1038-1959. <http://doi.org/10.6018/analesps.29.3.178511>
- Ato, M., & Vallejo, G. (2015). *Diseños de investigación en psicología*. Madrid: Pirámide.
- Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, 82(1), 60-70. http://doi.org/10.1207/s15327752jpa8201_11
- Beaton, D. ., Bombardier, C., Katz, J. ., & Wright, J. . (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemiology*, 54(12), 1204-1217. [http://doi.org/10.1016/S0895-4356\(01\)00407-3](http://doi.org/10.1016/S0895-4356(01)00407-3)

- Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 9(2), 78-84.
<http://doi.org/10.1027/1614-2241/a000057>
- Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy*, 17(3), 305-308.
[http://doi.org/10.1016/S0005-7894\(86\)80060-0](http://doi.org/10.1016/S0005-7894(86)80060-0)
- Cleary, P. D., & McNeil, B. J. (1988). Patient satisfaction as an indicator of quality care. *Inquiry*, 25, 25-36.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1s Ed.). Hillsdale, NJ, England: Lawrence Erlbaum Associates.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: L. Erlbaum Associates.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.
<http://doi.org/10.1037/0033-2909.112.1.155>
- Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A)*, 20(5), 755-762.
<http://doi.org/10.1076/jcen.20.5.755.1132>
- Crawford, J. R., Howell, D. C., & Garthwaite, P. H. (1998). Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples t test. *Journal of Clinical and Experimental Neuropsychology*, 20(6), 898-905. <http://doi.org/10.1076/jcen.20.6.898.1112>
- Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*, 56(5), 395-407.
[http://doi.org/10.1016/S0895-4356\(03\)00044-1](http://doi.org/10.1016/S0895-4356(03)00044-1)

- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61(4), 532-574. <http://doi.org/10.1177/0013164401614002>
- Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, 7(4), 485-503. <http://doi.org/10.1037/1082-989X.7.4.485>
- De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, 9(1), 123-149. <http://doi.org/10.1146/annurev-clinpsy-050212-185617>
- De Vet, H. C. W., Ostelo, R. W. J. G., Terwee, C. B., Roer, N., Knol, D. L., Beckerman, H., ... Bouter, L. M. (2006). Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Quality of Life Research*, 16(1), 131-142. <http://doi.org/10.1007/s11136-006-9109-9>
- Fayers, P. M., & Machin, D. (2000). *Quality of Life: assessment, analysis and interpretation*. New York: Wiley and Sons Ltd.
- Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods*, 14(1), 43-53. <http://doi.org/10.1037/a0014699>
- Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532-538. <http://doi.org/10.1037/a0015808>
- Ferrer, R. (2014). *Significación clínica: falsos positivos y falsos negativos en la estimación del cambio individual*. Universidad Autónoma de Madrid, Madrid.

- Ferrer, R., & Pardo, A. (2014). Clinically meaningful change: False positives in the estimation of individual change. *Psychological Assessment*, 26(2), 370-383. <http://doi.org/10.1037/a0035419>
- Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, 79(2), 314-316. <http://doi.org/http://dx.doi.org/10.1037/0021-9010.79.2.314>
- Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: univariate and multivariate applications* (2nd ed). New York: Routledge.
- Guyatt, G. H., Juniper, E. F., Walter, S. D., Griffith, L. E., & Goldstein, R. S. (1998). Interpreting treatment effects in randomised trials. *BMJ*, 316(7132), 690-693. <http://doi.org/10.1136/bmj.316.7132.690>
- Hageman, W. J. J. M., & Arrindell, W. A. (1999). Establishing clinically significant change: increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy*, 37(12), 1169-1193. [http://doi.org/10.1016/S0005-7967\(99\)00032-7](http://doi.org/10.1016/S0005-7967(99)00032-7)
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36. <http://doi.org/10.1148/radiology.143.1.7063747>
- Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, 15(4), 336-352. [http://doi.org/10.1016/S0005-7894\(84\)80002-7](http://doi.org/10.1016/S0005-7894(84)80002-7)
- Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, 67(3), 300-307. <http://doi.org/10.1037/0022-006X.67.3.300>

- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12-19. <http://doi.org/10.1037/0022-006X.59.1.12>
- Kahn, R., Goldfarb, A. I., Pollack, M., & Peck, A. (1960). Brief objective measures for the determination of mental status in the aged. *American Journal of Psychiatry*, 117(4), 326-328.
- Kirk, R. E. (2013). *Experimental design: procedures for the behavioral sciences* (4th ed). Thousand Oaks: Sage Publications.
- Lambert, M. J., Hansen, N. B., & Bauer, S. (2008). Assessing the clinical significance of outcome results. En A. M. Nezu & C. M. Nezu (Eds.), *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions* (pp. 359-378). New York, NY, US: Oxford University Press.
- Lemieux, J., Beaton, D. E., Hogg-Johnson, S., Bordeleau, L. J., & Goodwin, P. J. (2007). Three methods for minimally important difference: no relationship was found with the net proportion of patients improving. *Journal of Clinical Epidemiology*, 60(5), 448-455. <http://doi.org/10.1016/j.jclinepi.2006.08.006>
- Leon, O. G., & Montero, I. (2003). *Métodos de investigación en psicología y educación* (3.^a ed.). Madrid: McGraw-Hill / Interamericana de España.
- Maassen, G. H. (2004). The standard error in the Jacobson and Truax Reliable Change Index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society*, 10(6), 888-893. <http://doi.org/10.1017/S1355617704106097>
- Martinovich, Z., Saunders, S., & Howard, K. (1996). Some Comments on «Assessing Clinical Significance». *Psychotherapy Research*, 6(2), 124-132. <http://doi.org/10.1080/10503309612331331648>

- McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111(2), 361-365. <http://doi.org/http://dx.doi.org/10.1037/0033-2909.111.2.361>
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156-166. <http://doi.org/10.1037/0033-2909.105.1.156>
- Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, 11(2), 364-386. <http://doi.org/10.1177/1094428106291059>
- Morris, T. R., Cho, C., Dilda, V., Shine, J. M., Naismith, S. L., Lewis, S. J. G., & Moore, S. T. (2012). A comparison of clinical and objective measures of freezing of gait in Parkinson's disease. *Parkinsonism & Related Disorders*, 18(5), 572-577. <http://doi.org/10.1016/j.parkreldis.2012.03.001>
- Norman, G. R., Sridhar, F. G., Guyatt, G. H., & Walter, S. D. (2001). Relation of distribution and anchor-based approaches in interpretation of changes in Health-Related Quality of Life. *Medical Care*, 39(10), 1039-1047. <http://doi.org/10.1097/00005650-200110000-00002>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3.^a ed.). New York: McGraw-Hill.
- Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review*, 21(3), 421-446. [http://doi.org/10.1016/S0272-7358\(99\)00058-6](http://doi.org/10.1016/S0272-7358(99)00058-6)
- Oltmanns, T. F., & Turkheimer, E. (2009). Person perception and personality pathology. *Current Directions in Psychological Science*, 18(1), 32-36. <http://doi.org/10.1111/j.1467-8721.2009.01601.x>

- Ottenbacher, K. J., Johnson, M. B., & Hojem, M. (1988). The significance of clinical change and clinical change of significance: Issues and methods. *American Journal of Occupational Therapy*, 42(3), 156-163. <http://doi.org/10.5014/ajot.42.3.156>
- Owens, J., Sangal, R. B., Sutton, V. K., Bakken, R., Allen, A. J., & Kelsey, D. (2009). Subjective and objective measures of sleep in children with attention-deficit/hyperactivity disorder. *Sleep Medicine*, 10(4), 446-456. <http://doi.org/10.1016/j.sleep.2008.03.013>
- Pardo, A., & Ferrer, R. (2013). Significación clínica: falsos positivos en la estimación del cambio individual. *Anales de Psicología*, 29(2). <http://doi.org/10.6018/analesps.29.2.139031>
- Pardo, A., Garrido, J., Ruiz, M. A., & San Martín, R. (2007). La interacción entre factores en el análisis de varianza: errores de interpretación. *Psicothema*, 19(2), 343-349.
- Pardo, A., & San Martín, R. (2015). *Análisis de datos en ciencias sociales y de la salud II* (2.^a ed.). Madrid: Síntesis.
- Payne, R. W., & Jones, H. G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology*, 13(2), 115-121. [http://doi.org/10.1002/1097-4679\(195704\)13:2<115::AID-JCLP2270130203>3.0.CO;2-1](http://doi.org/10.1002/1097-4679(195704)13:2<115::AID-JCLP2270130203>3.0.CO;2-1)
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: an integrated approach*. Hillsdale, N.J: Lawrence Erlbaum Associates.
- Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, 61(2), 102-109. <http://doi.org/10.1016/j.jclinepi.2007.03.012>
- Schmitt, J. S., & Di Fabio, R. P. (2004). Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *Journal of Clinical Epidemiology*, 57(10), 1008-1018. <http://doi.org/10.1016/j.jclinepi.2004.02.007>

- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shikiar, R., & Rentz, A. M. (2004). Satisfaction with medication: An overview of conceptual, methodologic, and regulatory issues. *Value in Health*, 7(2), 204-215. <http://doi.org/10.1111/j.1524-4733.2004.72252.x>
- Testa, M. A. (2000). Interpretation of quality-of-life outcomes: issues that affect magnitude and meaning. *Medical Care*, 38(9 Suppl), II166-74.
- Tingey, R., Lambert, M., Burlingame, G., & Hansen, N. (1996). Assessing Clinical Significance: Proposed Extensions to Method. *Psychotherapy Research*, 6(2), 109-123. <http://doi.org/10.1080/10503309612331331638>
- Vargha, A., & Delaney, H. D. (2000). A Critique and Improvement of the «CL» Common Language Effect Size Statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2), 101-132. <http://doi.org/10.2307/1165329>
- Vindras, P., Desmurget, M., & Baraduc, P. (2012). When one size does not fit all: A simple statistical method to deal with across-individual variations of effects. *PLoS ONE*, 7(6), e39059. <http://doi.org/10.1371/journal.pone.0039059>
- Winer, B. J., Brown, D. R., & Michels, K. . (1991). *Statistical Principles in Experimental Design* (3.^a ed.). New York, NY: McGraw-Hill.
- Wright, J. G. (2000). Evaluating the outcome of treatment. Shouldn't we be asking patients if they are better? *Journal of Clinical Epidemiology*, 53(6), 549-553. [http://doi.org/10.1016/S0895-4356\(99\)00225-5](http://doi.org/10.1016/S0895-4356(99)00225-5)
- Wyrwich, K. W., Tierney, W. M., & Wolinsky, F. D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in Health-Related Quality of Life. *Journal of Clinical Epidemiology*, 52(9), 861-873. [http://doi.org/10.1016/S0895-4356\(99\)00071-2](http://doi.org/10.1016/S0895-4356(99)00071-2)

Apéndice 1

**Equivalencias entre estadísticos basados en los
centros de la distribución y estadísticos basados en
los individuos**

Apéndice 1a – Diseño pre-post de grupo único

Porcentaje de mejorías individuales en función del tamaño del efecto $d = M_{\text{dif}} / S_{\text{dif}}$

d	$R_{XY} = .5$							$R_{XY} = .7$							$R_{XY} = .9$						
	$As = -3, Cr = 18$	$As = -2, Cr = 9$	$As = -1, Cr = 2$	$As = 0, Cr = 0$	$As = 1, Cr = 2$	$As = 2, Cr = 9$	$As = 3, Cr = 18$	$As = -3, Cr = 18$	$As = -2, Cr = 9$	$As = -1, Cr = 2$	$As = 0, Cr = 0$	$As = 1, Cr = 2$	$As = 2, Cr = 9$	$As = 3, Cr = 18$	$As = -3, Cr = 18$	$As = -2, Cr = 9$	$As = -1, Cr = 2$	$As = 0, Cr = 0$	$As = 1, Cr = 2$	$As = 2, Cr = 9$	$As = 3, Cr = 18$
.0	3.0	3.0	3.6	5.0	5.9	5.7	5.8	1.9	2.1	3.0	5.0	6.2	6.2	6.2	0.6	0.9	2.1	5.1	6.5	6.4	6.6
.3	4.9	5.7	7.3	8.8	9.2	8.5	8.0	3.5	4.5	6.6	8.9	9.3	8.9	8.3	1.7	3.3	5.9	9.1	9.7	8.9	8.7
.6	8.5	11.0	13.2	15.0	13.9	12.4	11.3	7.3	10.1	13.3	14.9	13.9	12.6	11.6	4.7	9.5	12.8	15.1	14.1	12.7	12.0
.9	15.0	19.3	21.7	23.1	20.5	18.4	15.7	15.3	20.5	23.0	23.0	20.6	18.2	16.1	16.8	21.2	23.2	23.1	20.3	17.8	15.9
1.2	29.5	32.6	34.0	32.9	29.6	26.7	23.1	33.4	35.1	35.1	33.1	29.0	25.8	22.3	39.0	37.2	36.1	33.0	28.4	25.3	21.7
1.5	50.8	48.2	47.0	44.5	41.0	37.8	33.2	54.6	50.9	48.3	44.5	39.5	36.7	31.8	57.5	53.0	49.8	44.0	39.1	34.6	30.5
1.8	66.7	62.4	59.7	56.1	53.6	52.1	49.8	68.7	64.0	60.6	56.5	52.1	50.1	46.0	70.1	65.5	61.4	56.3	51.0	47.5	42.8
2.1	77.3	74.0	70.9	67.4	66.5	67.7	71.5	77.7	74.7	71.2	67.4	65.3	65.4	66.9	78.5	75.2	71.9	67.6	64.2	63.2	61.2
2.4	84.3	82.2	79.5	77.6	77.9	80.8	85.0	84.2	82.3	80.0	77.2	77.5	80.2	85.4	84.2	82.4	80.0	77.6	76.8	78.8	84.3
2.7	88.8	87.8	86.1	85.2	86.9	89.6	91.5	88.7	87.7	86.1	85.1	86.8	90.0	93.3	88.5	87.4	86.0	85.2	87.5	90.8	95.3
3.0	92.0	91.7	90.9	91.4	92.8	94.6	95.3	91.7	91.3	90.9	91.1	93.5	95.5	96.4	91.3	91.0	90.6	91.1	94.3	96.7	98.3
3.3	94.3	94.4	94.1	95.1	96.5	97.0	97.2	93.9	93.9	93.9	95.0	97.3	97.8	98.1	93.4	93.7	93.7	95.1	98.1	99.1	99.5
3.6	95.8	96.1	96.3	97.5	98.3	98.5	98.3	95.5	95.7	96.1	97.5	98.9	99.1	99.0	95.1	95.5	95.7	97.4	99.4	99.7	99.7

Se consideran como mejorías los individuos con un valor de $SID > 1.64$ (excepto en las condiciones con $d = 0$, donde se aplica $SID > 1.96$).

Cada casilla informa del valor promedio de las 500 muestras simuladas en cada condición con $n = 100$.

Estos porcentajes han sido calculados para un efecto expresado en la métrica de las diferencias pre-post. Si solo se conoce el tamaño del efecto expresado en la métrica de las puntuaciones pre ($d_2 = M_{\text{dif}} / S_{\text{pre}}$), este valor puede transformarse mediante $d = d_2[2(1 - R_{XY})]^{1/2}$ y posteriormente buscar el valor de d en la tabla (ver Introducción).

Apéndice 1b

Diseño de grupos aleatorios, medidos solo en post

Probabilidad de Superioridad (*PS*) del grupo tratado sobre el control

en función del tamaño del efecto $d = \frac{M_{tto} - M_{ctrl}}{S}$

d	$As = -3,$ $Cr = 18$	$As = -2,$ $Cr = 9$	$As = -1,$ $Cr = 2$	$As = 0,$ $Cr = 0$	$As = 1,$ $Cr = 2$	$As = 2,$ $Cr = 9$	$As = 3,$ $Cr = 18$
.0	.501	.503	.500	.502	.500	.500	.498
.3	.649	.609	.592	.584	.592	.609	.651
.6	.751	.709	.679	.662	.679	.707	.750
.9	.819	.784	.756	.737	.757	.784	.821
1.2	.870	.845	.820	.803	.819	.846	.871
1.5	.905	.891	.868	.856	.870	.888	.905
1.8	.929	.922	.907	.898	.908	.920	.930
2.1	.945	.942	.937	.930	.936	.943	.948
2.4	.959	.959	.957	.955	.955	.959	.958
2.7	.969	.970	.970	.972	.970	.970	.969
3.0	.976	.979	.980	.983	.980	.979	.976
3.3	.981	.984	.986	.990	.986	.984	.982
3.6	.986	.988	.991	.995	.991	.988	.986

La probabilidad de superioridad se calcula de forma no paramétrica como

$$PS = [n_{tto} n_{ctrl} + n_{ctrl} (n_{ctrl} + 1) / 2 - S_{ctr}] / (n_{tto} n_{ctr})$$

Cada casilla informa del valor promedio de las 500 muestras simuladas en cada condición con $n = 100$. Si solo se conoce el tamaño del efecto expresado como correlación biserial-puntual (R_{bp}) entre la variable dicotómica (que define los dos grupos) y la cuantitativa (en la cual se están comparando los grupos), esta correlación se puede transformar en d mediante la expresión

$$d = 2 R_{bp} / \sqrt{1 - R_{bp}^2}$$

Posteriormente puede buscarse el valor de d en la tabla (ver Introducción).

Apéndice 1c – Diseño pre-post con grupos aleatorios

Porcentaje de neto de mejorías individuales en función del tamaño del efecto $\hat{\omega}^2 = \frac{gl_{interacción} (F_{interacción} - 1)}{gl_{interacción} (F_{interacción} - 1) + N}$

d_{tto}	$\hat{\omega}^2$	$R_{XY} = .5$							$R_{XY} = .7$							$R_{XY} = .9$						
		$As = -3,$ $Cr = 18$	$As = -2,$ $Cr = 9$	$As = -1,$ $Cr = 2$	$As = 0,$ $Cr = 0$	$As = 1,$ $Cr = 2$	$As = 2,$ $Cr = 9$	$As = 3,$ $Cr = 18$	$As = -3,$ $Cr = 18$	$As = -2,$ $Cr = 9$	$As = -1,$ $Cr = 2$	$As = 0,$ $Cr = 0$	$As = 1,$ $Cr = 2$	$As = 2,$ $Cr = 9$	$As = 3,$ $Cr = 18$	$As = -3,$ $Cr = 18$	$As = -2,$ $Cr = 9$	$As = -1,$ $Cr = 2$	$As = 0,$ $Cr = 0$	$As = 1,$ $Cr = 2$	$As = 2,$ $Cr = 9$	$As = 3,$ $Cr = 18$
.0	.00	.0	-.1	-.3	.3	-.1	-.2	-.2	.0	.1	-.1	.2	.2	-.2	.2	.2	.1	.2	.2	.0	-.1	-.1
.3	.01	2.1	2.6	3.1	3.6	3.2	2.6	2.0	2.1	2.5	3.3	3.8	2.7	2.1	2.0	1.5	1.8	2.6	3.7	2.6	1.9	1.8
.6	.04	5.1	5.9	7.4	8.1	6.6	5.5	4.9	4.4	5.6	7.3	8.3	6.5	5.1	4.3	3.4	5.2	7.2	8.5	5.7	4.5	3.8
.9	.09	8.8	11.2	14.0	14.5	11.7	9.7	8.2	8.6	11.1	14.4	14.4	10.9	8.8	7.5	6.8	12.1	15.1	14.6	10.1	8.2	6.8
1.2	.15	16.1	20.4	23.2	22.2	18.2	15.3	12.6	16.9	22.2	24.4	22.5	17.3	14.3	12.0	17.9	24.0	25.6	22.9	16.4	13.4	11.1
1.5	.22	29.1	33.9	35.0	32.4	27.2	23.6	19.7	34.8	36.8	36.4	32.3	25.7	22.2	18.6	41.9	39.8	38.3	32.5	24.4	20.4	16.6
1.8	.28	52.0	49.3	48.1	43.2	38.0	34.6	30.1	56.4	52.1	49.8	43.6	35.9	32.4	27.6	61.0	55.1	52.5	43.6	34.6	30.3	25.1
2.1	.35	68.5	64.2	61.1	55.8	51.0	49.5	45.9	70.9	66.3	62.7	55.4	49.0	46.0	41.1	73.1	68.4	64.4	55.7	46.8	42.5	37.3
2.4	.42	79.4	75.5	71.9	67.2	64.2	64.8	68.2	80.7	77.1	73.2	66.9	62.4	61.0	63.0	82.2	78.3	74.9	66.6	60.1	57.6	55.9
2.7	.47	86.1	83.9	81.0	77.4	75.6	78.5	81.9	87.4	84.9	82.6	77.1	74.1	76.1	81.1	88.3	85.9	83.4	76.7	72.7	74.2	78.7
3.0	.53	90.8	90.0	88.0	85.3	85.1	86.9	88.7	91.6	90.8	88.6	84.8	83.8	86.4	89.2	92.6	91.3	89.4	85.1	83.3	86.1	90.1
3.3	.57	94.3	94.1	92.7	90.8	90.9	92.1	92.7	94.7	94.4	93.4	91.1	90.8	92.1	92.9	95.3	95.3	93.9	90.8	90.4	92.5	93.4
3.6	.62	96.7	96.4	96.0	94.9	94.6	94.6	94.6	97.2	96.9	96.5	94.9	94.3	94.5	94.6	98.1	97.7	96.9	94.6	94.3	94.8	94.8

Se consideran como mejorías los individuos con un valor de $SID > 1.96$. Al porcentaje de casos que cumplen esta condición se le resta el porcentaje de casos que empeoran significativamente ($SID < -1.96$). Después se resta el porcentaje neto del grupo control al del grupo tratado. Cada casilla informa del valor promedio de las 500 muestras simuladas en cada condición con $n = 100$. Para facilitar la interpretación se incluye d_{tto} . Este estadístico es la diferencia media pre-post expresada en la métrica de las diferencias para el grupo tratado. El valor esperado de d_{ctr} es cero en todas las condiciones. Si sólo se conoce el tamaño del efecto expresado mediante d_{PPC} , éste estadístico puede transformarse en $\hat{\omega}^2$ mediante la expresión:

$$\hat{\omega}^2 = \frac{d_{PPC}^2 n_{tto} n_{ctrl} S_{pre}^2 - (n_{tto} + n_{ctrl}) S_{dif}^2}{d_{PPC}^2 n_{tto} n_{ctrl} S_{pre}^2 + (n_{tto} + n_{ctrl}) (4n_{tto} - 1) S_{dif}^2}$$

Apéndice 2

Código de MatLab utilizado para las simulaciones

Apéndice 2a

Simulación de un diseño pre-post de grupo único. Código de MatLab

```
% Eduardo Estrada - eduardo.estrada@uam.es
%
% effects size range = [0:.3:3.6] (13 effect sizes)
%
% [data] = sampleGenLoop(n, nRepl, skewn, rxy, write)
% n = sample size
% nRepl = number of replicas for each effect size
% skewn = skewness of distributions
% kurt = kurtosis of distributions
% rxy = pearson correlation between measures pre and post treatment
% write = [true/false] write data into a text file

function [data] = sampleGenLoop(n, nRepl, skewn, kurt, rxy, write)

ESs = 0 : .3 : 3.6; % Effect sizes definition
nES = length(ESs); % Length of ES vector (number of samples to create)
nStats = 11;

COR = [1,rxy;rxy,1]; % correlation matrix between x and y

% initiate empty matrix for output global database
totalNrows = nRepl* n * nES;
data = zeros(totalNrows,nStats);
writeInRowGlob = 1;

for repl=1:nRepl

    dataForReplLoop = zeros(n * nES,nStats);
    writeInRowLocal = 1;

    for ES = 1:nES

        X = pearsrnd(0,1,skewn,kurt,n,2); % generate matrix X of random
                                         values
                                         % 2 columns and 'nsuj' rows
        MAT = X*(cholcov(COR)); % Cholesky-like covariance decomposition of
                                'COR'
                                % multiplying X

        varx = MAT(:,1); % re-scale x to defined metric

        vary = MAT(:,2); % re-scale y to defined metric

        difYX = vary-varx; % calculate vector of differences between y and x
                           % for each row

        DESVDIF = std(difYX); % standard deviation of the differences

        EStoApply = ESs(ES)*DESVdif;

        vary1 = vary + repmat(EStoApply,n,1);

        skewx = repmat((skewness(varx)), n, 1);
        skewy = repmat((skewness(vary)), n, 1);
```

```

    skewdifYX = repmat((skewness(difYX)), n, 1);
    kurtx = repmat((kurtosis(varx)), n, 1);
    kurty = repmat((kurtosis(vary)), n, 1);
    kurtdifYX = repmat((kurtosis(difYX)), n, 1);

    nrepl = zeros(n,1);

    dataForESloop = [nrepl+repl, repmat(ESs(ES),n,1), varx, vary1,...
        repmat(EStoApply, n, 1), skewx, skewy, skewdifYX,...
        kurtx, kurty, kurtdifYX];

    dataRows = size(dataForESloop,1); % numer of rows in current database

    % write current database in output global database
    dataForReplLoop( writeInRowLocal:writeInRowLocal + dataRows-1, : ) =
dataForESloop;

    % dataForESloop from next loop will be written right after
dataForESloop from this loop
    writeInRowLocal = writeInRowLocal+dataRows;

end % for ES = 1:nES

dataRows = size(dataForReplLoop,1); % numer of rows in current database

% write current database in output global database
data( writeInRowGlob:writeInRowGlob + dataRows-1, : ) = dataForReplLoop;

% dataForESloop from next loop will be written right after dataForESloop
from this loop
    writeInRowGlob = writeInRowGlob+dataRows;

end % for repl=1:replicas

filename = strcat(num2str(nRepl),'loop', 'N', num2str(n),...
    'skewn', num2str(skewn), 'Rxy', num2str(rxy), '.txt');

if write == true
    disp('now writing data on file')
    disp(filename)
    disp(' ')
    dlmwrite(filename,data,'delimiter','\t','precision','%8.3f');
else
    disp(filename)
    disp('will not be written on hard disk')
    disp(' ')
end % if write == true

end % function sampleGenLoop

-----

```

```

nRepl = 500;
n1 = 25;
n2 = 50;
n3 = 100;

skewn_array = [-3 -2 -1 0 1 2 3];
rxy_array = [.5 .7 .9];

nConditions = length(skewn_array)*length(rxy_array);

%individualData = cell(nConditions,2,3);
groupData = cell(nConditions,2,3);

%% program to generate individual data
tic;
condition = 1;

for skewn = skewn_array
    if (skewn == 0)
        kurt = 3;
        disp([skewn kurt])
    else
        kurt = skewn^2 + abs(skewn)*4;
        disp([skewn kurt])
    end

    for rxy = rxy_array
        datanamelist = genvarname({...
strcat('dataN1skn',num2str(skewn),'krt',num2str(kurt),'r',num2str(rxy*10)),...
strcat('dataN2skn',num2str(skewn),'krt',num2str(kurt),'r',num2str(rxy*10)),...
strcat('dataN3skn',num2str(skewn),'krt',num2str(kurt),'r',num2str(rxy*10))});

        individualData{condition,1,1} = datanamelist{1};
        individualData{condition,2,1} = sampleGenLoop(n1, nRepl, skewn, kurt,
rxy, true);

        individualData{condition,1,2} = datanamelist{2};
        individualData{condition,2,2} = sampleGenLoop(n2, nRepl, skewn, kurt,
rxy, true);

        individualData{condition,1,3} = datanamelist{3};
        individualData{condition,2,3} = sampleGenLoop(n3, nRepl, skewn, kurt,
rxy, true);

        condition = condition + 1;
    end % for r = rxy_array
end; % for skewn in skewn_array

save('individualData', 'individualData', '-v7');
indivTimer = toc/60

```


Apéndice 2b

Simulación de un diseño de grupos aleatorios. Código de MatLab

```
% Eduardo Estrada - eduardo.estrada@uam.es
%
% Control and experimental group, post-treatment measure
% effects size range = [0:.3:3.6] (13 effect sizes)
%
% [filename, data] = sampleGenLoop(n, nRepl, skewn, write)
% n = sample size
% nRepl = number of replicas for each effect size
% skewn = skewness of distributions
% kurt = kurtosis of distributions
% value
% write = [true/false] write data into a text file

%function [data] = sampleGenLoop(n, nRepl, skewn, kurt, effectDispersion,
    write)
function [data] = sampleGenLoop(n, nRepl, skewn, kurt, write)

ESs = 0 : .3 : 3.6; % Effect sizes definition
nES = length(ESs); % Length of ES vector (number of samples to create)
nStats = 9;

% initiate empty matrix for output global database
totalNrows = nRepl * n * nES;
data = zeros(totalNrows, nStats);
writeInRowGlob = 1;

for repl=1:nRepl

    dataForReplLoop = zeros(n * nES, nStats);
    writeInRowLocal = 1;

    for ES = 1:nES

        % Control group data

        vary1_Ctr = pearsrnd(0,1,skewn,kurt,n,1); % generate matrix X_Ctr
                                                    % of random values
                                                    % 2 columns and 'nsuj'
                                                    rows

        skewy_Ctr = repmat((skewness(vary1_Ctr)), n, 1);
        kurty_Ctr = repmat((kurtosis(vary1_Ctr)), n, 1);

        % Experimental group

        %EStoApply = ESs(ES) + normrnd(0,effectDispersion);

        %vary1_Exp = pearsrnd(EStoApply,1,skewn,kurt,n,1);
        vary1_Exp = pearsrnd(ESs(ES),1,skewn,kurt,n,1);

        skewy_Exp = repmat((skewness(vary1_Exp)), n, 1);
        kurty_Exp = repmat((kurtosis(vary1_Exp)), n, 1);
```

```

% Data preparation

dataForESloop = [zeros(n,1)+repl, repmat(ESs(ES),n,1),...
    varyl_Ctr, varyl_Exp,...
    repmat(ESs(ES), n, 1),...
    skewy_Ctr, kurty_Ctr, skewy_Exp, kurty_Exp];

dataRows = size(dataForESloop,1); % numer of rows in current database

% write current database in output global database
dataForReplLoop( writeInRowLocal:writeInRowLocal + dataRows-1, : ) =
dataForESloop;

% dataForESloop from next loop will be written right after
% dataForESloop from this loop
writeInRowLocal = writeInRowLocal+dataRows;

end % for ES = 1:nES

dataRows = size(dataForReplLoop,1); % numer of rows in current database

% write current database in output global database
data( writeInRowGlob:writeInRowGlob + dataRows-1, : ) = dataForReplLoop;

% dataForESloop from next loop will be written right after dataForESloop
% from this loop
writeInRowGlob = writeInRowGlob+dataRows;

end % for repl=1:replicas

filename = strcat(num2str(nRepl),'loop', 'N', num2str(n),...
    'skewn', num2str(skewn), '.txt');

if write == true
    disp('now writing data on file')
    disp(filename)
    disp(' ')
    dlmwrite(filename,data,'delimiter','\t','precision','%8.3f');
else
    disp(filename)
    disp('will not be written on hard disk')
    disp(' ')
end % if write == true

end % function sampleGenLoop

-----

```

```

nRepl = 500;
n1 = 10;
n2 = 25;
n3 = 50;
n4 = 100;

%effectDispersion = .4;
saveRawData = false;
saveSummData = true;

skewn_array = [-3 -2 -1 0 1 2 3];

nConditions = length(skewn_array);

individualData = cell(nConditions,2,3);
groupData = cell(nConditions,2,3);

%% program to generate individual data
tic;
condition = 1;

for skewn = skewn_array
    if (skewn == 0)
        kurt = 3;
        disp([skewn kurt])
    else
        kurt = skewn^2 + abs(skewn)*4;
        disp([skewn kurt])
    end

    datanamelist = genvarname({...
        strcat('dataN1skn',num2str(skewn),'krt',num2str(kurt)),...
        strcat('dataN2skn',num2str(skewn),'krt',num2str(kurt)),...
        strcat('dataN3skn',num2str(skewn),'krt',num2str(kurt)),...
        strcat('dataN4skn',num2str(skewn),'krt',num2str(kurt))});

    individualData{condition,1,1} = datanamelist{1};
    individualData{condition,2,1} = sampleGenLoop(n1, nRepl, skewn, kurt,
saveRawData);

    individualData{condition,1,2} = datanamelist{2};
    individualData{condition,2,2} = sampleGenLoop(n2, nRepl, skewn, kurt,
saveRawData);

    individualData{condition,1,3} = datanamelist{3};
    individualData{condition,2,3} = sampleGenLoop(n3, nRepl, skewn, kurt,
saveRawData);

    individualData{condition,1,4} = datanamelist{4};
    individualData{condition,2,4} = sampleGenLoop(n4, nRepl, skewn, kurt,
saveRawData);

    condition = condition + 1;

end; % for skewn in skewn_array

indivTimer = toc/60

```


Apéndice 2c

Simulación de un diseño pre-post con grupos aleatorios. Código de MatLab

```
% Eduardo Estrada - eduardo.estrada@uam.es
%
% Control and experimental group
% effects size range = [0:.3:3.6] (13 effect sizes)
%
% [filename, data] = sampleGenLoop(n, nRepl, skewn, rxy, write)
% n = sample size
% nRepl = number of replicas for each effect size
% skewn = skewness of distributions
% kurt = kurtosis of distributions
% rxy = pearson correlation between measures pre and post treatment
% write = [true/false] write data into a text file

function [data] = sampleGenLoop(n, nRepl, skewn, kurt, rxy, write)

ESs = 0 : .3 : 3.6; % Effect sizes definition
nES = length(ESs); % Length of ES vector (number of samples to create)
nStats = 19;

COR = [1,rxy;rxy,1]; % correlation matrix between x and y

% initiate empty matrix for output global database
totalNrows = nRepl* n * nES;
data = zeros(totalNrows,nStats);
writeInRowGlob = 1;

for repl=1:nRepl

    dataForReplLoop = zeros(n * nES,nStats);
    writeInRowLocal = 1;

    for ES = 1:nES

        % Control group data

        X_Ctr = pearsrnd(0,1,skewn,kurt,n,2); % generate matrix X_Ctr of
                                                random values
                                                % 2 columns and 'nsuj' rows
        MAT_Ctr = X_Ctr *(cholcov(COR)); % Cholesky-like covariance
                                                decomposition of 'COR'
                                                % multiplying X_Ctr

        varx_Ctr = MAT_Ctr(:,1); % re-scale x to defined metric

        vary1_Ctr = MAT_Ctr(:,2); % re-scale y to defined metric

        difYX_Ctr = vary1_Ctr - varx_Ctr; % calculate vector of differences
                                                between y and x
                                                % for each row

        skewx_Ctr = repmat((skewness(varx_Ctr)), n, 1);
        skewy_Ctr = repmat((skewness(vary1_Ctr)), n, 1);
        skewdifYX_Ctr = repmat((skewness(difYX_Ctr)), n, 1);
        kurtx_Ctr = repmat((kurtosis(varx_Ctr)), n, 1);
        kurty_Ctr = repmat((kurtosis(vary1_Ctr)), n, 1);
```

```

kurtdifYX_Ctr = repmat((kurtosis(difYX_Ctr)), n, 1);

% Experimental group

X_Exp = pearsrnd(0,1,skewn,kurt,n,2); % generate matrix X_Ctr of
                                     random values
                                     % 2 columns and 'nsuj' rows
MAT_Exp = X_Exp * (cholcov(COR)); % Cholesky-like covariance
                                   decomposition of 'COR'
                                   % multiplying X_Exp

varx_Exp = MAT_Exp(:,1); % re-scale x to defined metric

vary_Exp = MAT_Exp(:,2); % re-scale y to defined metric

difYX_Exp = vary_Exp - varx_Exp; % calculate vector of differences
                                   between y and x
                                   % for each row

DESVDIF_Exp = std(difYX_Exp); % standard deviation of the differences

EStoApply_Exp = ESs(ES)*DESVDIF_Exp;

varyl_Exp = vary_Exp + repmat(EStoApply_Exp,n,1);

skewx_Exp = repmat((skewness(varx_Exp)), n, 1);
skewy_Exp = repmat((skewness(varyl_Exp)), n, 1);
skewdifYX_Exp = repmat((skewness(difYX_Exp)), n, 1);
kurtx_Exp = repmat((kurtosis(varx_Exp)), n, 1);
kurty_Exp = repmat((kurtosis(varyl_Exp)), n, 1);
kurtdifYX_Exp = repmat((kurtosis(difYX_Exp)), n, 1);

% Data preparation

%nsrepl = zeros(n,1);

dataForESloop = [zeros(n,1)+repl, repmat(ESs(ES),n,1),...
                 varx_Ctr, varyl_Ctr, varx_Exp, varyl_Exp...
                 repmat(EStoApply_Exp, n, 1),...
                 skewx_Ctr, skewy_Ctr, skewdifYX_Ctr,...
                 kurtx_Ctr, kurty_Ctr, kurtdifYX_Ctr,...
                 skewx_Exp, skewy_Exp, skewdifYX_Exp,...
                 kurtx_Exp, kurty_Exp, kurtdifYX_Exp];

dataRows = size(dataForESloop,1); % number of rows in current database

% write current database in output global database
dataForReplLoop( writeInRowLocal:writeInRowLocal + dataRows-1, : ) =
dataForESloop;

% dataForESloop from next loop will be written right after
% dataForESloop from this loop
writeInRowLocal = writeInRowLocal+dataRows;

end % for ES = 1:nES

dataRows = size(dataForReplLoop,1); % number of rows in current database

```

```

    % write current database in output global database
    data( writeInRowGlob:writeInRowGlob + dataRows-1, : ) = dataForReplLoop;

    % dataForESloop from next loop will be written right after dataForESloop
    from this loop
    writeInRowGlob = writeInRowGlob+dataRows;

end % for repl=1:replicas

filename = strcat(num2str(nRepl),'loop', 'N', num2str(n),...
                  'skewn', num2str(skewn), 'Rxy', num2str(rxy), '.txt');

if write == true
    disp('now writing data on file')
    disp(filename)
    disp(' ')
    dlmwrite(filename,data,'delimiter','\t','precision','%8.3f');
else
    disp(filename)
    disp('will not be written on hard disk')
    disp(' ')
end % if write == true

end % function sampleGenLoop

```

```

nRepl = 500;
n1 = 10;
n2 = 25;
n3 = 50;
n4 = 100;

saveRawData = false;
saveSummData = true;

skewn_array = [-3 -2 -1 0 1 2 3];
rxy_array = [.5 .7 .9];

nConditions = length(skewn_array)*length(rxy_array);

%individualData = cell(nConditions,2,3);
groupData = cell(nConditions,2,3);

%% program to generate individual data
tic;
condition = 1;

for skewn = skewn_array
    if (skewn == 0)
        kurt = 3;
        disp([skewn kurt])
    else
        kurt = skewn^2 + abs(skewn)*4;
        disp([skewn kurt])
    end

    for rxy = rxy_array
        datanamelist = genvarname({...

strcat('dataN1skn',num2str(skewn),'krt',num2str(kurt),'r',num2str(rxy*10)),...
strcat('dataN2skn',num2str(skewn),'krt',num2str(kurt),'r',num2str(rxy*10)),...
strcat('dataN3skn',num2str(skewn),'krt',num2str(kurt),'r',num2str(rxy*10)),...
strcat('dataN4skn',num2str(skewn),'krt',num2str(kurt),'r',num2str(rxy*10))});

        individualData{condition,1,1} = datanamelist{1};
        individualData{condition,2,1} = sampleGenLoop(n1, nRepl, skewn, kurt,
rxy, saveRawData);

        individualData{condition,1,2} = datanamelist{2};
        individualData{condition,2,2} = sampleGenLoop(n2, nRepl, skewn, kurt,
rxy, saveRawData);

        individualData{condition,1,3} = datanamelist{3};
        individualData{condition,2,3} = sampleGenLoop(n3, nRepl, skewn, kurt,
rxy, saveRawData);

        individualData{condition,1,4} = datanamelist{4};
        individualData{condition,2,4} = sampleGenLoop(n4, nRepl, skewn, kurt,
rxy, saveRawData);

        condition = condition + 1;

    end % for r = rxy_array
end; % for skewn in skewn_array

indivTimer = toc/60

```




Facultad de Psicología

Departamento de Psicología Social y

Metodología